F21	375	JAPIO
F22	244*	BIOTECHABS
F23	244*	BIOTECHDS
F24	218	AGRICOLA
F25	200	NLDB
F26	177	TOXLINE
F27	172	DPCI
F28	151	PATOSWO
F29	123	JICST-EPLUS
F30	118	TULSA
F31	113	AQUASCI
F32	80	GENBANK
F33	45	BIOBUSINESS
F34	45*	INPADOC
F35	41	OCEAN
F36	40	AIDSLINE
F37	40	EMBAL
F38	37*	CEABA
F39	31	ANABSTR
F40	30	CJELSEVIER
F41	30	PIRA
F42	28	DRUGU
F43	22*	FSTA
F44	19*	CIN
F45	17	CEN
F46	17*	RAPRA
F47	15	DDFU
F48	14	CONFSCI
F49	12	PHIN
F50	8	PNI
F51	6	CROPU
F52	6	PAPERCHEM2
F53	4	JPNEWS
F54	3	HEALSAFE
F55	3*	APIPAT
F56	1	DDFB
F57	1	DRUGB
F58	1	TULSA2

W/ Dianne Reese

=> file f1-f58

COST IN U.S. DOLLARS	SINCE FILE ENTRY	TOTAL SESSION
FULL ESTIMATED COST	3.15	4.20

FILE 'USPATFULL' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'IFIPAT' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 IFI/Plenum Data Corporation (IFI)

FILE 'CAPLUS' ENTERED AT 08:45:14 ON 28 APR 1997
USE IS SUBJECT TO THE TERMS OF YOUR CUSTOMER AGREEMENT
COPYRIGHT (C) 1997 AMERICAN CHEMICAL SOCIETY (ACS)

FILE 'EUROPATFULL' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (c) 1997 WILA Verlag Muenchen (WILA)

FILE 'WPIDS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 DERWENT INFORMATION LTD

FILE 'WPINDEX' ACCESS NOT AUTHORIZED

FILE 'BIOSIS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 BIOSIS(R)

FILE 'SCISEARCH' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Institute for Scientific Information (ISI) (R)

FILE 'MEDLINE' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'EMBASE' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Elsevier Science B.V. All rights reserved.

FILE 'TOXLIT' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'LIFESCI' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Cambridge Scientific Abstracts (CSA)

FILE 'CJACS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 American Chemical Society (ACS)

FILE 'DISSABS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 University Microfilms, Inc.

FILE 'NTIS' ENTERED AT 08:45:14 ON 28 APR 1997
Compiled and distributed by the National Technical Information Service
of the Department of Commerce of the United States of America.
All rights reserved. (1997) (NTIS)

FILE 'PROMT' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Information Access Company. All rights reserved.

FILE 'INPADOC' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT 1997 (C) European Patent Office, Vienna (EPO)

FILE 'OCEAN' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Cambridge Scientific Abstracts (CSA)

FILE 'AIDSLINE' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'EMBAL' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Elsevier Science B.V. All rights reserved.

FILE 'CEABA' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (c) 1997 DECHEMA, ROY SOC CHEM, FIZ CHEMIE

FILE 'ANABSTR' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (c) 1997 THE ROYAL SOCIETY OF CHEMISTRY (RSC)

FILE 'CJELSEVIER' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Elsevier Science Publishers B.V. (ELSEVIER)

FILE 'PIRA' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Pira International

FILE 'DRUGU' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 DERWENT INFORMATION LTD

FILE 'FSTA' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 International Food Information Service

FILE 'CIN' ENTERED AT 08:45:14 ON 28 APR 1997
USE IS SUBJECT TO THE TERMS OF YOUR CUSTOMER AGREEMENT
COPYRIGHT (C) 1997 American Chemical Society (ACS)

FILE 'CEN' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 American Chemical Society (ACS)

FILE 'RAPRA' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 RAPRA Technology Ltd.

FILE 'DDFU' ACCESS NOT AUTHORIZED

FILE 'CONFSCI' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Cambridge Scientific Abstracts (CSA)

FILE 'PHIN' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'DGENE' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 DERWENT INFORMATION LTD

FILE 'CABA' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 CAB INTERNATIONAL (CABI)

FILE 'PATOSEP' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (c) 1997 WILA Verlag Muenchen (WILA)

FILE 'CANCERLIT' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'JAPIO' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Japanese Patent Office (JPO) and Japan Patent
Information Organization (Japio)

FILE 'BIOTECHABS' ACCESS NOT AUTHORIZED

FILE 'BIOTECHDS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 DERWENT INFORMATION LTD

FILE 'AGRICOLA' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'NLDB' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Information Access Company. All rights reserved.

FILE 'TOXLINE' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'DPCI' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 DERWENT INFORMATION LTD

FILE 'PATOSWO' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (c) 1997 WILA Verlag Muenchen (WILA)

FILE 'JICST-EPLUS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Japan Science and Technology Corporation (JST)

FILE 'TULSA' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 The University of Tulsa (UTULSA)

FILE 'AQUASCI' ENTERED AT 08:45:14 ON 28 APR 1997
(c) 1997 FAO (on behalf of the ASFA Advisory Board) All rights reserved.

FILE 'GENBANK' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'BIOBUSINESS' ENTERED AT 08:45:14 ON 28 APR 1997
COPYRIGHT (C) 1997 Biological Abstracts, Inc. (BIOSIS)

FILE 'PNI' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'CROPU' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'PAPERCHEM2' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'JPNEWS' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'HEALSAFE' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'APIPAT' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'DDFB' ACCESS NOT AUTHORIZED

FILE 'DRUGB' ENTERED AT 08:45:14 ON 28 APR 1997

FILE 'TULSA2' ENTERED AT 08:45:14 ON 28 APR 1997

=> s (SBH or sequenc###(p)array#)

L2 14206 FILE USPATFULL
L3 4819 FILE IFIPAT
L4 2696 FILE CAPLUS
L5 2571 FILE EUROPATFULL
L6 2439 FILE WPIDS
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L7 2048 FILE BIOSIS
L8 1913 FILE SCISEARCH
L9 1884 FILE MEDLINE
L10 1575 FILE EMBASE
L11 1154 FILE TOXLIT
L12 981 FILE LIFESCI
L13 915 FILE CJACS

```
L14         746 FILE DISSABS
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L15         728 FILE NTIS
L16         511 FILE PROMT
L17         481 FILE DGENE
L18         456 FILE CABA
L19         400 FILE PATOSEP
L20         399 FILE CANCERLIT
L21         375 FILE JAPIO
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L22         244 FILE BIOTECHDS
L23         218 FILE AGRICOLA
L24         200 FILE NLDB

L25         177 FILE TOXLINE
L26         172 FILE DPCI
L27         151 FILE PATOSWO
L28         123 FILE JICST-EPLUS
L29         118 FILE TULSA
L30         113 FILE AQUASCI

L31          80 FILE GENBANK
L32          45 FILE BIOBUSINESS
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L33          45 FILE INPADOC
L34          41 FILE OCEAN
L35          40 FILE AIDSLINE
L36          40 FILE EMBAL
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L37          37 FILE CEABA
L38          31 FILE ANABSTR
L39          30 FILE CJELSEVIER
L40          30 FILE PIRA
L41          28 FILE DRUGU
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L42          22 FILE FSTA
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L43          19 FILE CIN
L44          17 FILE CEN
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
```

```
L45          17 FILE RAPRA
L46          14 FILE CONFSCI
L47          12 FILE PHIN
L48           8 FILE PNI
L49           6 FILE CROPU
L50           6 FILE PAPERCHEM2
L51           4 FILE JPNEWS
L52           3 FILE HEALSAFE
PROXIMITY OPERATOR LEVEL NOT CONSISTENT WITH
FIELD CODE - 'AND' OPERATOR ASSUMED 'SEQUENC###(P)ARRAY#'
L53           3 FILE APIPAT
L54           1 FILE DRUGB
L55           1 FILE TULSA2

TOTAL FOR ALL FILES
L56       43393 (SBH OR SEQUENC###(P) ARRAY#)

=> s l56 and (mismatch or single base)

L57         519 FILE USPATFULL
L58          25 FILE IFIPAT
L59          37 FILE CAPLUS
L60         123 FILE EUROPATFULL

L61           9 FILE WPIDS
L62          27 FILE BIOSIS
L63          23 FILE SCISEARCH
L64          26 FILE MEDLINE
L65          21 FILE EMBASE
L66          24 FILE TOXLIT
L67          11 FILE LIFESCI
L68          58 FILE CJACS
L69           6 FILE DISSABS
L70           2 FILE NTIS
L71           2 FILE PROMT
L72          21 FILE DGENE
L73           6 FILE CABA
L74           1 FILE PATOSEP
L75           5 FILE CANCERLIT
L76           0 FILE JAPIO
L77           9 FILE BIOTECHDS
L78           2 FILE AGRICOLA
L79           3 FILE NLDB
L80           3 FILE TOXLINE
L81           0 FILE DPCI
L82           0 FILE PATOSWO
L83           1 FILE JICST-EPLUS
```

```
L84           0 FILE TULSA
L85           0 FILE AQUASCI
L86           0 FILE GENBANK
L87           0 FILE BIOBUSINESS
L88           0 FILE INPADOC
L89           0 FILE OCEAN
L90           0 FILE AIDSLINE
L91           0 FILE EMBAL
L92           0 FILE CEABA
L93           0 FILE ANABSTR
L94           4 FILE CJELSEVIER
L95           0 FILE PIRA
L96           0 FILE DRUGU
L97           0 FILE FSTA
L98           0 FILE CIN
L99           1 FILE CEN
L100          0 FILE RAPRA
L101          0 FILE CONFSCI
L102          0 FILE PHIN
L103          0 FILE PNI
L104          0 FILE CROPU
L105          0 FILE PAPERCHEM2
L106          0 FILE JPNEWS
L107          0 FILE HEALSAFE
L108          0 FILE APIPAT
L109          0 FILE DRUGB
L110          0 FILE TULSA2

TOTAL FOR ALL FILES
L111        969 L56 AND (MISMATCH OR SINGLE BASE)

=> s l111 and (intensit### or signal# or label#)

SEARCH ENDED BY USER

=> s l111 and (intensit### or signal# or label# or pixel#)


L112        482 FILE USPATFULL

L113         19 FILE IFIPAT
L114          4 FILE CAPLUS

L115        118 FILE EUROPATFULL

L116          4 FILE WPIDS
L117          5 FILE BIOSIS
```

L118	4	FILE	SCISEARCH
L119	4	FILE	MEDLINE
L120	4	FILE	EMBASE
L121	3	FILE	TOXLIT
L122	2	FILE	LIFESCI
L123	48	FILE	CJACS
L124	2	FILE	DISSABS
L125	1	FILE	NTIS
L126	2	FILE	PROMT
L127	2	FILE	DGENE
L128	1	FILE	CABA
L129	0	FILE	PATOSEP
L130	0	FILE	CANCERLIT
L131	0	FILE	JAPIO
L132	4	FILE	BIOTECHDS
L133	1	FILE	AGRICOLA
L134	2	FILE	NLDB
L135	0	FILE	TOXLINE
L136	0	FILE	DPCI
L137	0	FILE	PATOSWO
L138	1	FILE	JICST-EPLUS
L139	0	FILE	TULSA
L140	0	FILE	AQUASCI
L141	0	FILE	GENBANK
L142	0	FILE	BIOBUSINESS
L143	0	FILE	INPADOC
L144	0	FILE	OCEAN
L145	0	FILE	AIDSLINE
L146	0	FILE	EMBAL
L147	0	FILE	CEABA
L148	0	FILE	ANABSTR
L149	2	FILE	CJELSEVIER
L150	0	FILE	PIRA
L151	0	FILE	DRUGU
L152	0	FILE	FSTA
L153	0	FILE	CIN
L154	1	FILE	CEN
L155	0	FILE	RAPRA
L156	0	FILE	CONFSCI
L157	0	FILE	PHIN
L158	0	FILE	PNI
L159	0	FILE	CROPU
L160	0	FILE	PAPERCHEM2
L161	0	FILE	JPNEWS
L162	0	FILE	HEALSAFE
L163	0	FILE	APIPAT
L164	0	FILE	DRUGB

L165 0 FILE TULSA2

TOTAL FOR ALL FILES
L166 716 L111 AND (INTENSIT### OR SIGNAL# OR LABEL# OR PIXEL#)

=> s l166 and (computer# or program# or software#)

L167 348 FILE USPATFULL
L168 4 FILE IFIPAT
L169 0 FILE CAPLUS

L170 101 FILE EUROPATFULL
L171 0 FILE WPIDS
L172 0 FILE BIOSIS
L173 0 FILE SCISEARCH
L174 0 FILE MEDLINE
L175 0 FILE EMBASE
L176 0 FILE TOXLIT
L177 0 FILE LIFESCI
L178 36 FILE CJACS
L179 1 FILE DISSABS
L180 1 FILE NTIS

L181 1 FILE PROMT
L182 0 FILE DGENE
L183 0 FILE CABA
L184 0 FILE PATOSEP
L185 0 FILE CANCERLIT
L186 0 FILE JAPIO
L187 0 FILE BIOTECHDS
L188 0 FILE AGRICOLA

L189 1 FILE NLDB
L190 0 FILE TOXLINE
L191 0 FILE DPCI
L192 0 FILE PATOSWO
L193 1 FILE JICST-EPLUS
L194 0 FILE TULSA
L195 0 FILE AQUASCI
L196 0 FILE GENBANK
L197 0 FILE BIOBUSINESS
L198 0 FILE INPADOC
L199 0 FILE OCEAN
L200 0 FILE AIDSLINE
L201 0 FILE EMBAL
L202 0 FILE CEABA
L203 0 FILE ANABSTR

```
L204          2 FILE CJELSEVIER
L205          0 FILE PIRA
L206          0 FILE DRUGU
L207          0 FILE FSTA
L208          0 FILE CIN
L209          1 FILE CEN
L210          0 FILE RAPRA
L211          0 FILE CONFSCI
L212          0 FILE PHIN
L213          0 FILE PNI
L214          0 FILE CROPU
L215          0 FILE PAPERCHEM2
L216          0 FILE JPNEWS
L217          0 FILE HEALSAFE
L218          0 FILE APIPAT
L219          0 FILE DRUGB
L220          0 FILE TULSA2

TOTAL FOR ALL FILES
L221        497 L166 AND (COMPUTER# OR PROGRAM# OR SOFTWARE#)

=> s l221 and (ratio# or comparison#)

L222        279 FILE USPATFULL
L223          1 FILE IFIPAT
L224          0 FILE CAPLUS
L225         78 FILE EUROPATFULL
L226          0 FILE WPIDS
L227          0 FILE BIOSIS
L228          0 FILE SCISEARCH
L229          0 FILE MEDLINE
L230          0 FILE EMBASE
L231          0 FILE TOXLIT
L232          0 FILE LIFESCI
L233         33 FILE CJACS
L234          1 FILE DISSABS
L235          0 FILE NTIS
L236          0 FILE PROMT
L237          0 FILE DGENE
L238          0 FILE CABA
L239          0 FILE PATOSEP
L240          0 FILE CANCERLIT
L241          0 FILE JAPIO
L242          0 FILE BIOTECHDS
L243          0 FILE AGRICOLA
L244          0 FILE NLDB
L245          0 FILE TOXLINE
```

```
L246          0 FILE DPCI
L247          0 FILE PATOSWO
L248          0 FILE JICST-EPLUS
L249          0 FILE TULSA
L250          0 FILE AQUASCI
L251          0 FILE GENBANK
L252          0 FILE BIOBUSINESS
L253          0 FILE INPADOC
L254          0 FILE OCEAN
L255          0 FILE AIDSLINE
L256          0 FILE EMBAL
L257          0 FILE CEABA
L258          0 FILE ANABSTR
L259          2 FILE CJELSEVIER
L260          0 FILE PIRA
L261          0 FILE DRUGU
L262          0 FILE FSTA
L263          0 FILE CIN
L264          0 FILE CEN
L265          0 FILE RAPRA
L266          0 FILE CONFSCI
L267          0 FILE PHIN
L268          0 FILE PNI
L269          0 FILE CROPU
L270          0 FILE PAPERCHEM2
L271          0 FILE JPNEWS
L272          0 FILE HEALSAFE
L273          0 FILE APIPAT
L274          0 FILE DRUGB
L275          0 FILE TULSA2

TOTAL FOR ALL FILES
L276        394 L221 AND (RATIO# OR COMPARISON#)

=> s l276 and (oligonucleotide# or DNA or gene# or nucleic acid#)


L277         50 FILE USPATFULL
L278          1 FILE IFIPAT

L279          0 FILE CAPLUS

L280         12 FILE EUROPATFULL

L281          0 FILE WPIDS
L282          0 FILE BIOSIS
L283          0 FILE SCISEARCH
```

L284 0 FILE MEDLINE

L285 0 FILE EMBASE
L286 0 FILE TOXLIT

L287 0 FILE LIFESCI
L288 29 FILE CJACS
L289 1 FILE DISSABS
L290 0 FILE NTIS
L291 0 FILE PROMT
L292 0 FILE DGENE
L293 0 FILE CABA
L294 0 FILE PATOSEP

L295 0 FILE CANCERLIT
L296 0 FILE JAPIO
L297 0 FILE BIOTECHDS
L298 0 FILE AGRICOLA
L299 0 FILE NLDB

L300 0 FILE TOXLINE
L301 0 FILE DPCI
L302 0 FILE PATOSWO
L303 0 FILE JICST-EPLUS
L304 0 FILE TULSA
L305 0 FILE AQUASCI
L306 0 FILE GENBANK
L307 0 FILE BIOBUSINESS
L308 0 FILE INPADOC
L309 0 FILE OCEAN
L310 0 FILE AIDSLINE
L311 0 FILE EMBAL
L312 0 FILE CEABA
L313 0 FILE ANABSTR
L314 0 FILE CJELSEVIER
L315 0 FILE PIRA
L316 0 FILE DRUGU
L317 0 FILE FSTA
L318 0 FILE CIN
L319 0 FILE CEN
L320 0 FILE RAPRA
L321 0 FILE CONFSCI
L322 0 FILE PHIN
L323 0 FILE PNI
L324 0 FILE CROPU
L325 0 FILE PAPERCHEM2

```
L326            0 FILE JPNEWS
L327            0 FILE HEALSAFE
L328            0 FILE APIPAT
L329            0 FILE DRUGB
L330            0 FILE TULSA2

TOTAL FOR ALL FILES
L331           93 L276 AND (OLIGONUCLEOTIDE# OR DNA OR GENE# OR NUCLEIC ACID
                  #)

=> duplicate remove

ENTER L# LIST OR (END):l331

DUPLICATE IS NOT AVAILABLE IN 'DGENE, DPCI, GENBANK'. ANSWERS FROM THESE FILES

WILL BE CONSIDERED UNIQUE
DUPLICATE PREFERENCE IS 'USPATFULL, IFIPAT, EUROPATFULL, CJACS, DISSABS'
KEEP DUPLICATES FROM MORE THAN ONE FILE?  Y/(N):n


PROCESSING COMPLETED FOR L331
L332            92 DUPLICATE REMOVE L331 (1 DUPLICATE REMOVED)

=> d l332 1-92 bib ab kwic
```

L332 ANSWER 6 OF 92 USPATFULL
AN 97:14574 USPATFULL
TI Methods for sorting polynucleotides using ***oligonucleotide*** tags
IN Brenner, Sydney, Cambridge, England
PA Spectragen, Inc., Hayward, CA, United States (U.S. corporation)
PI US 5604097 970218
AI US 94-358810 941219 (8)
RLI Continuation-in-part of Ser. No. US 94-322348, filed on 13 Oct 1994, now abandoned
DT Utility
EXNAM Primary Examiner: Chambers, Jasemine C.; Assistant Examiner: Priebe, Scott D.
LREP Macevicz, Stephen C.
CLMN Number of Claims: 31
ECL Exemplary Claim: 1
DRWN 7 Drawing Figure(s); 6 Drawing Page(s)
LN.CNT 1834
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB The invention provides a method of tracking, identifying, and/or sorting classes or subpopulations of molecules by the use of ***oligonucleotide*** tags. ***Oligonucleotide*** tags of the invention each consist of a plurality of subunits 3 to 6 nucleotides in length selected from a minimally cross-hybridizing set. A subunit of a minimally cross-hybridizing set forms a duplex or triplex having two or more mismatches with the complement of any other subunit of the same set. The number of ***oligonucleotide*** tags available in a particular embodiment depends on the number of subunits per tag and on the length of the subunit. An important aspect of the invention is the use of the ***oligonucleotide*** tags for sorting polynucleotides by specifically hybridizing tags attached to the polynucleotides to their complements on solid phase supports. This embodiment provides a readily automated system for manipulating and sorting polynucleotides, particularly useful in large-scale parallel operations, such as large-scale ***DNA*** sequencing, mRNA fingerprinting, and the like, wherein many target polynucleotides or many segments of a single target polynucleotide are sequenced simultaneously.
TI Methods for sorting polynucleotides using ***oligonucleotide*** tags

tags for specific hybridization to the CPG microparticles.

DETD . . . e.g. comprising Hamamatsu model 9403-02 photomultipliers, a Stanford Research Systems model SR445 amplifier and model SR430 multichannel scaler, and digital ***computer*** , e.g. a 486-based ***computer*** . The ***computer*** generates a two dimensional map of the slide which registers the positions of the microparticles.

DETD . . . on the attached microparticles undergo 20 cycles of probe ligation, washing, detection, cleavage, and washing, in accordance with the preferred ***single*** ***base*** sequencing methodology described below. Within each detection step, the scanning system records the fluorescent emission corresponding the base identified at. . .

DETD where TAMRA, FAM, ROX, and JOE are spectrally resolvable fluorescent ***labels*** attached by way of Aminolinker II (all being available from Applied Biosystems, Inc., Foster City, Calif.); the bold faced nucleotides. . . four nucleotides, A, C, G, T. TAMRA (tetramethylrhodamine), FAM (fluorescein), ROX (rhodamine X), and JOE (2',7'-dimethoxy-4',5'-dichlorofluorescein) and their attachment to ***oligonucleotides*** is also described in Fung et al, U.S. Pat. No. 4,855,225.

DETD . . . target polynucleotide ends as follows: the probes are incubated for 60 minutes at 16.degree. C. with 200 units of T4 ***DNA*** ligase and the anchored target polynucleotide in T4 ***DNA*** ligase buffer; after washing, the target polynucleotide is then incubated with 100 units T4 polynucleotide kinase in the manufacturer's recommended. . . 30 minutes at 37.degree. C., washed, and again incubated for 30 minutes at 16.degree. C. with 200 units of T4 ***DNA*** ligase and the anchored target polynucleotide in T4 ***DNA*** ligase buffer. Washing is accomplished by successively flowing volumes of wash buffer over the slide, e.g. TE, disclosed in Sambrook. . . above). After the cycle of ligation-phosphorylation-ligation and a final washing, the attached microparticles are scanned for the presence of fluorescent ***label*** , the positions and characteristics of which are recorded by the scanning system. The labeled target polynucleotide, i.e. the ligated complex,. . .

DETD Exemplary ***computer*** ***program*** for generating minimally cross hybridizing sets

DETD ---

Program minxh

```
c
c
 integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
 dimension nbase(6)
c
c
```

```
 write(*,*)`ENTER SUBUNIT LENGTH`
 read(*,100)nsub
100 format(i1)
 open(1,file=`sub4.dat`,form=`formatted`,status=`new`)
c
c
  nset=0
  do 7000 m1=1,3
   do 7000 m2=1,3
    do 7000.  .  .  with mismatches .ge. ndiff
c                  in matrix mset2 starting at
c                  position 2.
c                  Next transfer contents
c                  of mset2 into mset1 and
c                  start
c                    ***comparisons***   again this time
c                  starting with subunit 3.
c                  Continue until all subunits
c                  undergo the   ***comparisons***  .
c
c
  npass=0
c
c
1700  continue
   kk=npass+2
   npass=npass+1
c
c
   do 1500 m=npass+2,jj
    n=0
    do 1600 j=1,nsub
     if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
 2     mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
 2     mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
      n=n+1
      endif
1600  continue
.  .  .
```

DETD

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(iii) NUMBER OF SEQUENCES: 16

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 38 nucleotides

(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:
GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA38
(2) INFORMATION FOR SEQ ID NO: 2:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:
AATTCGGATGATGCATGCATCGACCC26
(2) INFORMATION FOR SEQ ID NO: 3:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:
TCGAGTCATCCGAT14
(2) INFORMATION FOR SEQ ID NO: 4:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:
ATCGGATGACATCAAC16
(2) INFORMATION FOR SEQ ID NO: 5:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:
CTAGTCGACCA11
(2) INFORMATION FOR SEQ ID NO: 6:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:
NRRGATCYNNN11
(2) INFORMATION FOR SEQ ID NO: 7:
(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 22 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:
GGGTCGATGCATGCATCATCCG22
(2) INFORMATION FOR SEQ ID NO: 8:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:
ATCGGATGAC10
(2) INFORMATION FOR SEQ ID NO: 9:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:
ATCNNNNNAC10
(2) INFORMATION FOR SEQ ID NO: 10:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:
TCGAGTNNNNNGAT14
(2) INFORMATION FOR SEQ ID NO: 11:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:
ATCGGATGACATCAAC16
(2) INFORMATION FOR SEQ ID NO: 12:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 nucleotides
(B) TYPE: ***nucleic*** ***acid***
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:
NNNAGTTGATGTCATCCGAT20
(2) INFORMATION FOR SEQ ID NO: 13:

AN 464067 EUROPATFULL ED 19970307 EW 9703 FS PS
TIEN SOLID PHASE DIAGNOSIS OF MEDICAL CONDITIONS.
TIDE FESTPHASENDIAGNOSE VON MEDIZINISCHEN KONDITIONEN.
TIFR DIAGNOSTIC EN PHASE SOLIDE DE CONDITIONS MEDICALES.
IN UHLEN, Mathias, Kvarnbogatan 30, S-752 39 Uppsala, SE
PA CEMU BIOTEKNIK AB, Banergatan 21, S-752 37 Uppsala, SE
PAN 1082941
AG Dzieglewska, Hanna Eva et al, Frank B. Dehn & Co., European Patent Attorneys, 179 Queen Victoria Street, London EC4V 4EL, GB
AGN 73231
OS EPB1997005 EP 0464067 B1 970115
SO Wila-EPS-1997-H03-T1
DT Patent
LA Anmeldung in Englisch; Veroeffentlichung in Englisch
DS R AT; R BE; R CH; R DE; R DK; R ES; R FR; R GB; R IT; R LI; R LU; R NL; R SE
PIT EPB1 EUROPAEISCHE PATENTSCHRIFT (Internationale Anmeldung)
PI EP 464067 B1 970115
OD 920108
AI EP 90-904803 900315
PRAI GB 89-6641 890322
GB 89-6642 890322
RLI WO 90-EP454 900315 INTAKZ
WO 9011369 901004 INTPNR
REP EP 192168 A WO 86-05815 A
REN Nucleic Acid Research, Volume 16, No. 23, December 1988, IRL Press Ltd, (Oxford, GB), A.-C. SYVANEN et al.: "Quantification of Polymerase Chain Reaction Products by Affinity-Based Hybrid Collection", pages 11327-11338 CHEMICAL ABSTRACTS, Volume 111, No. 23, 4 December 1989, (Columbus, Ohio, US), see page 170 abstract 210118f, &
US-A-200876 (United States Dept. of Health and Human Services) 15 February 1989 Journal of Cellular Biochemistry, No. 13, Part E, M. UHLEN et al.: "Approches to Solid-Phase DNA Technology using PCR", page 310, Abstract WH 250
DETDEN This invention relates to a

AN 97:5078 CJACS

SO Analytical Chemistry, (1997), 69(8), 1510-1517. CODEN: ANCHAM. ISSN: 0003-2700

TI Multiplexed ***DNA*** Sequencing and Diagnostics by Hybridization with Enriched Stable Isotope ***Labels***

AU (1) Arlinghaus, Heinrich F. (*); (2) Kwoka, Margaret N.; (3) Guo, Xiao-Qin; (4) Jacobson, K. Bruce

CS (1,2,3,4) Atom Sciences, Inc., Oak Ridge, Tennessee 37830
(1,2,3,4) Health Science Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831

AB A new DNA diagnostic and sequencing system has been developed that uses time-of-flight resonance ionization mass spectrometry (TOF-RIMS) to provide a rapid method of analyzing stable isotope-labeled oligonucleotides in form 1 sequencing by hybridization (SBH). With form 1, the DNA is immobilized on a nylon membrane and enriched isotope-labeled individual oligonucleotide probes are free to seek out complementary DNAs during hybridization. The major advantage of this new approach is that multiple oligonucleotides can be labeled with different enriched isotopes and can all be simultaneously hybridized to the genosensor matrix. The probes can then be simultaneously detected with TOF-RIMS with high selectivity, sensitivity, and efficiency. By using isotopically enriched tin labels, up to 10 labeled oligonucleotides could be examined in a single hybridization to the DNA matrix. Greater numbers of labels are available if rare earth isotopes are employed. In the present study, matrices containing three different DNAs were prepared and simultaneously hybridized with two different probes under a variety of conditions. The results show that DNAs, immobilized on nylon surfaces, can be specifically hybridized to probes labeled with different enriched tin isotopes. Discrimination between complementary and noncomplementary sites of better than 100 was obtained in multiplexed samples. This new SBH method, which employs stable isotopic labels to locate target DNAs and TOF-RIMS to detect the labels, will be a very versatile and extensive multiplexing method.

TI Multiplexed ***DNA*** Sequencing and Diagnostics by Hybridization with Enriched Stable Isotope ***Labels***

TX TX(1) of 38. Sequencing by hybridization (***SBH***) promises to produce large amounts of sequence information, due to the parallel nature of the analysis, and is particularly well suited for genome diagnostics, sequencing cDNAs or partial sequencing of clones, ***DNA*** and RNA sequencing, ***gene*** polymorphism studies, and identification of expressed ***genes*** . ***SBH*** consists of hybridizing an ***oligonucleotide*** of known

sequence to a ***DNA*** on a solid surface whose sequence is being sought. ***SBH*** relies on the specific base-pairing rules in duplex ***DNA*** to infer sequence information by the detection of hybridized ***DNA*** . The main variables in ***SBH*** are the length and composition of the ***oligonucleotide*** , the attachment of the ***DNA*** or the ***oligonucleotide*** to the solid surface, the method of labeling and detection of the hybridization, and the conditions for hybridization. There are two main formats for ***SBH*** : form 1 and form 2. With form 1, the ***DNA*** is attached to the solid surface and labeled individual ***oligonucleotide*** probes are free to seek out the DNAs during hybridization.1-3.Footnote. .Footnote. .Footnote. With form 2, individual ***oligonucleotide*** probes are attached to the surface and labeled DNAs that have been sheared are free to seek out the ***oligonucleotides*** .4-8.Footnote. .Footnote. .Footnote. .Footnote. .Footnote. Form 1 is useful for assessing a large number of DNAs for the presence of a number of short defined sequences. Form 2 is useful for sequencing a smaller number of ***DNA*** fragments.

TX(2) of 38. This paper reports a new approach to ***label*** and detect ***oligonucleotides*** in ***SBH*** form 1.9.Footnote. Typically, large DNAs are attached to nylon membranes and hybridization/detection is performed by using radioactive phosphorus (32P) for labeling ***oligonucleotides*** and autoradiography for detection. 2,3 However, the use of radioactive materials has certain limitations, poses health hazards, and increases disposal. . . not very quantitative, has low spatial resolution, and requires long exposure time. Fluorescence labeling is another technique often employed in ***DNA*** studies. However, difficulties are encountered using fluorescence ***labels*** in ***SBH*** form 1 because of nonspecific binding and substrate fluorescence background. Employing nonradioactive enriched stable isotopes as ***labels*** and time-of-flight resonance ionization mass spectrometry (TOF-RIMS) as a detection method may allow sequencing of large numbers of DNAs faster. . . is possible with any other technique.10-14.Footnote. .Footnote. .Footnote. .Footnote. .Footnote. The major advantage of this new approach is that multiple ***oligonucleotides*** can be labeled with different enriched isotopes which can be simultaneously hybridized to immobilized ***DNA*** on a genosensor matrix. TOF-RIMS can then be used to detect them simultaneously with very high selectivity, sensitivity, and efficiency. By using isotopically enriched tin ***labels*** , up to 10 labeled ***oligonucleotides*** could be examined in a single hybridization to the ***DNA*** matrix. Furthermore, since these 10 isotopes of

tin do not decay, the labeled ***oligonucleotides*** may be stored indefinitely until time for their use. Ten enriched tin isotopes are available (Isotec, Inc., Miamisburg, OH) with. . . (>90% for most of them).15.Footnote. In addition to tin, it is possible to use rare earth and other elements for ***labels*** , thus tremendously expanding multiplexing possibilities.

TX(3) of 38. The new ***SBH*** method, employing stable isotopic labeling of probe ***oligonucleotides*** and detection by TOF-RIMS, will be a very extensive multiplexing method. It not only offers a significant increase in analysis. . . Additional experiments have evaluated the position effects of mismatched pairs as well. To evaluate such complexity, utilizing multiple enriched isotope ***labels*** will be advantageous so that the competition among several oligomers for hybridization at common sites may be studied.

TX(4) of 38. In addition to studying hybridization processes, multiplexing ***SBH*** form 1 procedures would be effective for production, organization, and storage of genome and cDNA libraries3,17.Footnote. as well as for screening genomic ***DNA*** from tens of thousands of individuals for several genetic mutations. This technique also includes, among its possibilities, error checking on ***DNA*** sequence data in combination with an alternative technique.3,18.Footnote. Again, the use of multiple isotopes would greatly expedite the error checking procedure. ***SBH*** could also be applied to supplement electrophoresis data so that compressions and other areas of difficulty could be analyzed by. . .

TX(5) of 38. RIMS . . . times higher than for the other elements in the sample.19.Footnote. Since the isotope shifts of most elements are small in ***comparison*** to the bandwidth of the RI lasers used in our experiments (7-12 GHz), all isotopes of a chosen element will. . .

TX(6) of 38. RI is also extremely sensitive and efficient. The ***intensity*** of modern pulsed dye lasers is sufficient to saturate both the bound-bound transitions and the ionization step, thereby assuring near-unit. . .

TX(7) of 38. RI . . . background can always be determined by detuning the laser wavelength by a few tenths of a nanometer and measuring the ***signal*** under the same experimental conditions.

TX(8) of 38. Much . . . method has not been studied as thoroughly

as SIRIMP and is often less quantitative because of shot-to-shot fluctuations in laser ***intensity*** and changes of the surface morphology with consecutive laser shots. LARIMP must continue to be refined so that the faster. . .

TX(9) of 38. EXPERIMENTAL . . . gun for charge compensation, an excimer laser system for sample atomization, a resonance postionization laser (repetition rate 30 Hz), a ***computer*** -controlled (x, y, z, .vphi.) sample manipulator, a high-resolution video imaging system for sample observation, and a TOF mass spectrometer detection. . . z with resolution of 1 .mu.m, and 360.degree. motion in .vphi.. The manipulator is positioned by stepper motors driven under ***computer*** control which can change positions at a speed of 10 000 steps/s. The sample holder is a carousel arrangement which. . . and positioning of this lens are accomplished by external manipulations. All the important instrument functions are controlled by an IBM-compatible ***computer*** .

TX(10) of 38. EXPERIMENTAL SECTION. Sample Preparation. M13mp18 single-stranded ***DNA*** was obtained from United States Biochemical Corp. (Cleveland, OH). Two plasmid vectors, pSP70 and pBR322, were obtained from Promega Corp. (Madison, WI). Tin-labeled 17-mer ***oligonucleotides*** were prepared at Oak Ridge National Laboratory (ORNL). Two ***oligonucleotide*** primers were prepared, M13(-20) and T7, with sequences of 5'-GTA AAA CGA CGG CCA GT-3' and 5'-AAT ACG ACT CAC TAT AG-3', respectively. The ***oligonucleotides*** , with hexylamine on the 5'-end, were purified and reacted with the N-hydroxysuccinimide (NHS) ester of triethylstannylpropionic acid (TESPA) that contained isotopically enriched tin (>90%).31.Footnote. Separate preparations of each of the ***oligonucleotides*** were labeled with different isotopes of tin. Maximum Strength Nytran nylon membranes were obtained from Schleicher & Schuell, Inc. (Keene,. . .

TX(12) of 38. EXPERIMENTAL . . . solution of 0.5 M NaOH/1.5 M NaCl in H2O and were allowed to air-dry in the clean hood. The target ***DNA*** , diluted in the NaOH/NaCl solution, was pipetted onto the nylon. Generally, 4 .mu.L of the 4 fmol/.mu.L target ***DNA*** solution was pipetted in 1 .mu.L increments. The spots were .apprx.3 mm in diameter. The samples were air-dried, and then. . . applied. The membranes were placed under a UV lamp at 254 nm and 120 000 .mu.J/cm 2 to bind the ***DNA*** to the membrane. Next, they were dried loosely covered for 1 h in an oven at 80 .degree.C.

TX(14) of 38. RESULTS AND DISCUSSION. Determination of Detection and Hybridization Parameters. Nylon samples, spotted with Sn

isotope-labeled ***oligonucleotide*** primers (17-mer), were used to measure experimental parameters for SIRIMP and LARIMP analysis. To obtain a ***signal*** using either of these techniques, the bond between Sn and the organic molecule must be broken so that free Sn atoms are available for detection. For LARIMP, which uses a laser beam to vaporize and atomize the Sn-labeled ***oligonucleotide*** , 193 nm (ArF) atomization laser light results in acceptable Sn fragmentation, but longer wavelengths do not provide atomization necessary for. . . unstable resonator optics, but there is generally no advantage in using laser spot sizes smaller than the size of the ***DNA*** spot. Practically, the physical size of the ***DNA*** spot will be the determining factor.

TX(15) of 38. RESULTS AND DISCUSSION. Since the ***signal*** is dependent on the laser ***intensity*** , it is important that the laser ***intensity*** be uniform throughout the measurement. For this reason, a very small laser spot size is not desirable. With extremely small spot sizes, the depth of focus is very small and the laser ***intensity*** fluctuates as a function of displacement from the focal point (the distance to the nylon surface may deviate up to. . . = 0.08 (193 nm light) and a displacement of 0.5 mm, the beam diameter is .apprx.40 .mu.m, and the beam ***intensity*** is reduced by a factor of 180. However, by selecting the proper focal length so that do = 25 .mu.m,. . . .theta. = 0.01 and the same displacement results in a beam diameter of 25.5 .mu.m, resulting in only a 4% ***intensity*** change. For this reason, we defocused the beam in these experiments.

TX(16) of 38. RESULTS AND DISCUSSION. We compared SIRIMP with LARIMP on tune-up samples. We achieved good Sn ***signals*** with both techniques, but the LARIMP ***signal*** was consistently higher than the SIRIMP ***signal*** (.apprx.50 times). This permitted quantifiable ***signal*** to be obtained averaging only five laser shots, while with SIRIMP, several hundred ion pulses were needed. Consequently, LARIMP was. . .

TX(17) of 38. RESULTS AND DISCUSSION. The general effect of hybridization and wash temperatures on the hybridization of ***DNA*** is well-known. At lower temperatures more mismatches occur, while at higher temperatures discrimination is increased.32.Footnote. A study was initiated to determine at which hybridization temperature the best discrimination between hybridized [118Sn-M13(-20) ***signal***] and nonhybridized sites (118Sn-T7 ***signal***) can be achieved. At the same time, the study was constructed to determine the discrimination differences for 2 and 20. . . for these two probes at 25 .degree.C with a hybridization time of 2 h. At 15 .degree.C, the noncomplementary 118Sn-T7

signal was much smaller than the ***signal*** from the complementary 118Sn-M13(-20) probe, but was still detectable. At 25 .degree.C, the discrimination between complementary and noncomplementary probes was nearly 700. Higher temperatures (35 and 45 .degree.C) resulted in the loss of ***signal*** for both the M13(-20) and the T7. Contrary to what was expected, the 20 h M13(-20) samples displayed lower 118Sn ***signals*** than the 2 h samples. One explanation could be that some of the bound ***DNA*** was released from the nylon over such a prolonged hybridization period. All remaining experiments were performed using a 2-h hybridization. . .

TX(18) of 38. RESULTS AND DISCUSSION. Studies were also performed in which the target ***DNA*** amount was kept at 17 fmol, but the amount of M13(-20) tin-labeled probe were varied in concentrations of 0.01, 0.05,. . . .mu.L was maintained for all of the samples. The samples prepared with probe concentrations of 10-1000 nM produced the highest ***signal*** for these conditions. The difference in the height of the ***signal*** between 10 and 1000 nM was small, but the ***signal*** was more consistent across the area of the spot at 100 nM than at 10 nM.

TX(19) of 38. RESULTS AND DISCUSSION. The M13mp18 target ***DNA*** amount was varied to determine the best concentration. The tin-labeled M13(-20) probe was maintained at 10 nM while the target ***DNA*** amounts of 0.1, 0.3, 0.7, 4, 8, 12, 17, 170, and 670 fmol were used. In order to achieve the desired amount of ***DNA*** on the nylon membrane, solution concentration and volume were varied. Concentrations of the target ***DNA*** solutions were varied between 4 and 85 fmol/.mu.L. The highest concentration of M13mp18 that was available was 4 fmol/.mu.L. For this reason, the samples with the two highest amounts of ***DNA*** (170 and 670 fmol) were made with ***DNA*** that was diluted in Tris-HCl buffer, rather than in the NaOH/NaCl mixture. The volume that was spotted varied between 1. . . The samples were analyzed with LARIMP using a .apprx.50 .mu.m diameter laser beam spot size. When 17 fmol of target ***DNA*** was immobilized on a .apprx.3 mm diameter area, the maximum ***signal*** was obtained. The data demonstrate that LARIMP is extremely sensitive. The amount of target ***DNA*** in the atomization laser beam area (.apprx.0.2 mm 2) is only 5 amol, and the amount of Sn-labeled ***DNA*** probe is even less in this area (hybridization efficiency <100%). These data and also the fact that during analysis only a fraction of the Sn-labeled ***DNA*** probe was removed by the atomization laser beam indicate that the LARIMP sensitivity is much smaller than 1 amol.

TX(20) of 38. RESULTS AND DISCUSSION. Multiplexing and Imaging.

Nylon matrices were prepared with three different types of ***DNA*** . A single-stranded 7249-mer control ***DNA*** (M13mp18), a double-stranded 2401 bp plasmid vector (pSP70), and a double-stranded 4300 bp ***DNA*** (pBR322) were all used as targets. M13mp18 has a binding site for M13(-20) but not for T7. Likewise, pSP70 has. . . After this, they were further diluted in the 0.5 M NaOH/1.5 M NaCl buffer which had been used for spotting ***DNA*** . For the spot that had a mixture of M13mp18 and pSP70, the solutions of the ***DNA*** were mixed before spotting. Concentrations for the individual DNAs after mixing were identical to solutions used to spot DNAs separately. Samples were hybridized with ***oligonucleotides*** that had been labeled separately with two different tin isotopes, 118Sn-M13(-20) and 120Sn-T7. These solutions were mixed at equal concentrations. . .

TX(21) of 38. RESULTS AND DISCUSSION. A total of four 3 mm spots of 17 fmol each of target ***DNA*** were applied with a pipet and bound to a nylon membrane; these included the three individual DNAs and one spot. . . 2..Figure. Another nylon piece, 1 .times. 1.5 cm, was prepared similarly, but with nine spots of the three types of ***DNA*** in an array, and was hybridized with 30 .mu.L of probe solution and analyzed as a whole piece. The resulting two-dimensional image of the isotope ***ratio*** is shown in Figure 3..Figure.

TX(22) of 38. RESULTS . . . and only M13mp18 (A) are located. Both the 118Sn and 120Sn peaks were present in the analysis of the mixed ***DNA*** spot, as expected.

TX(23) of 38. RESULTS AND DISCUSSION. The molar amounts for each ***DNA*** in the mixed spot were identical to the amounts of these DNAs in separate spots. The ***signal*** level of the mixed spot (A/B) 118Sn peak, which corresponds to the M13mp18 ***DNA*** , is .apprx.80% of the 118Sn ***signal*** on the M13mp18 single spot. The ***signal*** level of the 120Sn peak, which corresponds to the pSP70 ***DNA*** , is .apprx.70% of the pSP70 single spot. The ***signal*** levels of the 120Sn peaks are generally about half that of the 118Sn peaks. This may be because the M13mp18 ***DNA*** is single stranded and the pSP70 ***DNA*** is double stranded. The temperature used for denaturing (45 .degree.C) may not have been high enough to completely separate the ***DNA*** into single strands, which is necessary for hybridization.

TX(24) of 38. RESULTS . . . Sn-labeled primers used in our experiments. Based on the enrichment, for example, 118Sn-labeled M13(-20) was expected to have a 118Sn/120Sn ***ratio*** of 59:1. However, a ***ratio*** of only 25:1 was measured when a

concentrated solution of the labeled primer (10 pmol/.mu.L) was pipetted onto a nylon. . . be detected at the location of the immobilized M13mp18. For complex hybridization experiments, the enrichment factors for each of the ***labels*** can be measured and calculations can be performed to correct for the contamination error. Also, isotope ***ratio*** measurements can be used to discriminate ***signal*** from background.

TX(25) of 38. RESULTS . . . used in these experiments and not due to cross hybridization. To obtain a quantitative assessment of the cross hybridization, a ***ratio*** comparing the amount of complementary and noncomplementary hybridization (the discrimination) is used. Because of the isotopic contamination, the best means of obtaining numbers to determine the discrimination in this multiplexing experiment is to compare the ***signal*** for the isotope on the complementary spot to the ***signal*** on the spot for the pBR322, which was not complementary to either probe. The complementary/noncomplementary discrimination ***ratio*** is 130:1 for 118Sn-M13(-20) in the M13mp18 spot versus 118Sn-M13(-20) in the pBR322 spot.

TX(26) of 38. RESULTS . . . of a nine-spot matrix is shown at the top of Figure 3. A two-dimensional spectral LARIMP image of the isotope ***ratio*** of 118Sn/120Sn is seen in the bottom. In this image, it can be clearly seen that, in a multiplexed experiment, the two different probes hybridized specifically to the complementary ***DNA*** and were readily detected with LARIMP. In the image, the 118Sn isotope dominates in the M13mp18 spots (green) and the. . . dominates in the pSP70 spots (yellow), as expected. As seen previously in the four-dot samples, the 120Sn was lower in ***intensity*** than the 118Sn for the corresponding spots where they hybridize.

TX(27) of 38. RESULTS AND DISCUSSION. The ability to ***sequence*** large numbers of DNAs has been greatly improved by the preparation of membranes with ***DNA*** matrices using a robotic procedure developed by Drmanac. 18 This procedure can produce a ***DNA*** matrix, containing hundreds or thousands of DNAs on one membrane. Typically, these matrices are used with 32P-labeled ***oligonucleotides*** . In addition, Drmanac's laboratory at Hyseq, Inc., has developed extensive mathematical models for handling data and strategies for using ***oligonucleotides*** to determine ***sequence*** in the DNAs attached to the membranes. We used these preparation techniques to test the feasibility of using tin-labeled ***oligonucleotides*** for hybridization and LARIMP for detection on these robotically produced membranes. Two different test matrices were obtained, both

with several ***arrays*** containing 300 .mu.m diameter individual spots of ***DNA*** . The DNAs used by Drmanac to prepare the ***arrays*** , M13mp18, pSP70, and pBR322, were provided by Atom Sciences and were identical to those used in Figures 2 and 3.. . .

TX(28) of 38. RESULTS AND DISCUSSION. Figure 4 depicts a spectral and a three-dimensional LARIMP image of the 118Sn ***signal*** of the stainless steel pin-punched sample that was hybridized with 118Sn-M13(-20) alone. A 118Sn ***signal*** was detected on the M13mp18 spots, indicating that the 118Sn-M13(-20) hybridized to it, as expected. The 118Sn ***signal*** decreased proportionally with decreasing amounts of immobilized ***DNA*** , demonstrating the quantitative property of this technique. Repetitive sequences, which occur in eukaryotic ***DNA*** , are the primary source of errors in ***SBH*** and cannot be analyzed quantitatively with autoradiographic techniques. 3 Target ***DNA*** containing two identical sequences would likely hybridize to twice the amount of labeled probe, producing twice the LARIMP ***signal*** that a target ***DNA*** with a single matching sequence would produce. The indication that LARIMP can detect quantitative differences may mean that it can. . .

TX(29) of 38. RESULTS AND DISCUSSION. Figure 5 shows a two-dimensional LARIMP image of the 118Sn/120Sn ***ratio*** for the simultaneous multiplexing experiment of the stainless steel pin-punched matrix. The results were fully consistent with specific hybridization of. . . spots of pBR322, which should not have hybridized with either probe and the areas of the nylon membrane in which ***DNA*** was not immobilized are blue. This blue area displays either the natural ***ratio*** of 118Sn/120Sn (.apprx.0.75) or a tin ***signal*** that is below detectable levels. Most of the blue area observed in Figure 5 represents the absence of a detectable tin ***signal*** , and the level of noise can be seen in Figure 4.

TX(30) of 38. RESULTS . . . pipetting or robotic punching techniques and specifically hybridized with probes labeled with different enriched tin isotopes. The application of the ***DNA*** with pins afforded a much smaller and better defined spot than with the pipetted samples. Smaller spots will allow for. . . creating a cost savings for routine analysis due to reduced analysis time and reagents. The data indicate that this new ***SBH*** method, employing stable isotopic labeling and detection by LARIMP, will be a very versatile and extensive multiplexing method.

TX(31) of 38. CONCLUSION. A new ***DNA*** sequence analysis

system has been developed that uses TOF-RIMS to provide a rapid method of analyzing stable isotope-labeled ***oligonucleotides*** in ***SBH*** form 1. We demonstrated that DNAs can be immobilized on nylon surfaces and hybridized to probes labeled with specific enriched. . . complementary and noncomplementary sites of better than 100 was obtained in multiplexed samples. TOF-RIMS was adapted for detecting hybridized Sn-labeled ***oligonucleotides*** to immobilized ***DNA*** on nylon. RI can be used in two modes, LARIMP or SIRIMP. We have found that LARIMP offers quantitative analysis and is superior for the detection of Sn-labeled ***oligonucleotides*** on nylon.

TX(32) of 38. CONCLUSION. . . . and selectivity and detecting subattomole quantities with excellent lateral resolution. The major advantage of the TOF-RIMS instrument is that multiple ***oligonucleotides*** labeled with different enriched isotopes can be measured simultaneously. Tin has 10 isotopes and 2 of these have been used. . .

TX(33) of 38. CONCLUSION. To improve detection sensitivity further, ***oligonucleotide*** probes that are labeled with tin using the starburst dendrimer method developed by Tomalia34.Footnote. are being developed. Using this method, each ***oligonucleotide*** could be labeled with four, eight, or more tin atoms. This would substantially increase the sensitivity of the method. Also,. . . than the number of isotopes for each element. A set of simultaneous equations can be constructed from the measured isotope ***ratios*** to determine the background contamination and subtract it from the measured values. It may be possible to analyze human genomic ***DNA*** directly using starburst dendrimer and background subtraction technologies, if discrimination and repetitive sequences can be managed satisfactorily.

TX(34) of 38. CONCLUSION. In order to test the feasibility of using tin-labeled ***oligonucleotides*** for ***SBH*** form 1, a simple hybridization system and two labeled probes were used. In future experiments, however, discrimination will be determined. . . complementary pairs; each probe will be labeled with a different enriched isotope of tin. More complex hybrid formations, such as ***single*** ***base*** pair mismatches in different locations and varied A + T content will be included in the experiments. 32

TX(35) of 38. CONCLUSION. In the above described ***SBH*** procedure, the speed of ***DNA*** analysis is a strong function of the sensitivity, repetition rate, multiplexing extent, and miniaturization of the ***DNA*** -containing matrix. An excimer-based RI instrument can operate at 400 Hz. Triplicate

sampling of each ***pixel*** by the laser beam will allow an analysis rate of 133 ***pixels*** /s. Analysis for enriched stable isotopes of a 50 .times. 200 matrix (10 4 target sites with different DNAs) will require. . . sample translation. Furthermore, by using 21 enriched isotopes (8 Sn, 7 Nd, and 6 Gd isotopes), instead of 1, the ***DNA*** analysis speed would increase by a factor of 21. The speed of this technique far exceeds other detection methods currently. . .

TX(36) of 38. CONCLUSION. We conclude that enriched stable isotope-labeled ***DNA*** probes and LARIMP analysis has the potential to make a strong contribution to sequencing, diagnostics, mapping, and any other method. . .

TX(37) of 38. Acknowledgment. . . . Sciences, Inc. and M. J. Doktycz, Oak Ridge National Laboratory for helpful discussions and R. Drmanac, Hyseq, Inc. for attaching ***DNA*** to nylon membranes.

CP CP(2) of 5. Figure 2. Hybridization specificity with demonstration of simultaneous multiplexing. Hybridization was simultaneously carried out with two enriched Sn-labeled ***oligonucleotides*** . The top two drawings show an array design used to prepare multiplexing samples. (A) M13mp18 ***DNA*** , which contains a binding site for the 118Sn-M13(-20) ***oligonucleotide*** probe and not for the 120Sn-T7 probe. (B) pSP70 ***DNA*** , which has a binding site to the 120Sn-T7 probe and not for the 118Sn-M13(-20) probe. (C) pBR322 ***DNA*** , which does not have binding sites for either probe. In the bottom left, the results for the line scan of. . . array are shown. Enriched tin is not detected on the left side of the scan at the position where pBR322 ***DNA*** is bound. The peak on the right corresponds to 120Sn-T7, which binds to pSP70. Results for the right side of. . .

CP(3) of 5. Figure 3. Two-dimensional image of hybridized ***oligonucleotides*** labeled with two enriched tin isotopes. Hybridization was simultaneously carried out with both Sn-labeled ***oligonucleotides*** . The array design used to prepare nine-dot multiplexing samples is shown on top. The immobilized ***DNA*** and the Sn-labeled ***oligonucleotides*** are described in Figure 2. A two-dimensional image of the 118Sn/120Sn ***ratio*** on the nine-dot array is displayed at the bottom of the Figure. At the positions where M13mp18 (A in the top diagram) is bound, 118Sn dominates, giving a high 118Sn/120Sn ***ratio*** (green). Similarly, at positions where the pSP70 (B in the top diagram) is bound, 120Sn dominates, giving a low 118Sn/120Sn ***ratio*** (yellow). Natural isotopic ***ratios*** (118Sn/120Sn 0.75) were detected where the pBR322 ***DNA*** is located (C in the top

diagram) and over the remainder of the matrix.

CP(4) of 5. Figure . . . An array design used by Drmanac to prepare multiplexing samples is shown at the top of the figure. The immobilized ***DNA*** is described in Figure 2, and the 118Sn-M13(-20) probe was employed for hybridization. The numbers after the letters refer to. . . for pBR322, and 25 fmol/.mu.L for pSP70. Blank spots are marked with "xx". A two-dimensional spectral image of the 118Sn ***signal*** is displayed in the middle of the figure. The sample was hybridized with 118Sn-M13(-20). At the three positions where the M13mp18 is bound, 118Sn is detected at levels correlated to the amount of ***DNA*** bound to the nylon. No enriched Sn was observed at the pBR322 or pSP70 sites. The data shown in the middle is displayed three-dimensionally on the bottom of the Figure. The three peaks (in order of decreasing ***intensity***) correspond to decreasing M13mp18 sites concentrations of 84, 42, and 8 fmol/.mu.L.

CP(5) of 5. Figure 5. Hybridization specificity using three concentrations of three immobilized DNAs, robotically applied, and two different tin-labeled ***oligonucleotides*** . An array design used by Drmanac to prepare multiplexing samples is shown on top. The identities and concentrations of immobilized ***DNA*** are defined in Figure 4. Hybridization was carried out simultaneously using 118Sn-M13(-20) and 120Sn-T7. A two-dimensional image of the 118Sn/120Sn ***ratio*** in the ***DNA*** matrix prepared by Drmanac is displayed on the bottom. At the positions where the M13mp18 is bound, 118Sn is detected,. . . similarly, at the positions where the pSP70 is bound, 120Sn-T7 is detected. Enriched Sn is not detected where the pBR322 ***DNA*** is located. The level of ***signal*** is dependent upon the amount of ***DNA*** bound to the nylon.

L3

L332 ANSWER 14 OF 92 USPATFULL DUPLICATE 1
AN 96:85030 USPATFULL
TI Oligoprobe designstation: a computerized method for designing optimal ***DNA*** probes
IN Mitsuhashi, Masato, Irvine, CA, United States
Cooper, Allan J., Bellvue, WA, United States
Waterman, Michael S., Culver City, CA, United States
Pevzner, Pavel A., State College, PA, United States
PA Hitachi Chemical Research Center, Inc., Irvine, CA, United States (U.S. corporation)
PI US 5556749 960917
AI US 92-975526 921112 (7)
DT Utility
EXNAM Primary Examiner: Jones, W. Gray; Assistant Examiner: Tran, Paul L.
LREP Wagner & Middlebrook
CLMN Number of Claims: 97
ECL Exemplary Claim: 52
DRWN 160 Drawing Figure(s); 156 Drawing Page(s)
LN.CNT 2530
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB There is disclosed herein an invention which relates to the fields of genetic engineering, microbiology, and ***computer*** science, that allows a user, whether they be a molecular biologist or a clinical diagnostician, to calculate and design extremely specific ***oligonucleotide*** probes for ***DNA*** and mRNA hybridization procedures. The probes designed with this invention may be used for medical diagnostic kits, ***DNA*** identification, and potentially continuous monitoring of metabolic processes in human beings. The key features design ***oligonucleotide*** probes based on the GenBank database of ***DNA*** and mRNA sequences and examine candidate probes for specificity or commonality with respect to a user-selected experimental preparation. Two models are available: a ***Mismatch*** Model, that employs hashing and continuous seed filtration, and an H-Site Model, that analyzes candidate probes for their binding specificity relative to some known set of mRNA or ***DNA*** sequences. The preferred embodiment of this computerized design tool is written in the Borland.RTM. C++ language and runs under Microsoft.RTM. Windows.TM. on IBM.RTM. compatible personal ***computers*** .
TI Oligoprobe designstation: a computerized method for designing optimal ***DNA*** probes
AB There is disclosed herein an invention which relates to the fields of genetic engineering, microbiology, and ***computer***

science, that allows a user, whether they be a molecular biologist or a clinical diagnostician, to calculate and design extremely specific ***oligonucleotide*** probes for ***DNA*** and mRNA hybridization procedures. The probes designed with this invention may be used for medical diagnostic kits, ***DNA*** identification, and potentially continuous monitoring of metabolic processes in human beings. The key features design ***oligonucleotide*** probes based on the GenBank database of ***DNA*** and mRNA sequences and examine candidate probes for specificity or commonality with respect to a user-selected experimental preparation. Two models are available: a ***Mismatch*** Model, that employs hashing and continuous seed filtration, and an H-Site Model, that analyzes candidate probes for their binding specificity relative to some known set of mRNA or ***DNA*** sequences. The preferred embodiment of this computerized design tool is written in the Borland.RTM. C++ language and runs under Microsoft.RTM. Windows.TM. on IBM.RTM. compatible personal ***computers*** .

SUMM . . . three sheets of microfiche with a

PATENT APPLICATION - PATENTANMELDUNG - DEMANDE DE BREVET

AN 721016 EUROPATFULL UP 19970408 EW 9628 FS OS STA R

TIEN ***Nucleic*** ***acid*** library ***arrays*** , methods for synthesizing them and methods for ***sequencing*** and sample screening using them.

TIDE Immobilisierte Nukleinsaeure-Banken, Verfahren fuer ihre Herstellung und ihre Verwendung in Sequenzierungs- und Screeningsverfahren.

TIFR Banque d'acides nucleiques immobilisees, procedes pour leur fabrication et procedes de sequencage et de screening les utilisant.

IN Lockhart, David J., 480 Oakgrove Drive 205, Santa Clara, California 95054, US;
Chee, Mark S., 3199 Waverley Street, Palo Alto, California, 94306, US;
Vetter, Dirk, Husserlstr. 14, D-79110 Freiburg, DE;
Diggelmann, Martin, Dorfgasse 68, CH-4435 Niederdorf, CH

PA AFFYMAX TECHNOLOGIES N.V., De Ruyterkade 62, Willemstad, Curacao, AN

PAN 1316841

AG Bizley, Richard Edward et al, Hepworth, Lawrence, Bryer & Bizley Merlin House Falconry Court Baker's Lane, Epping Essex CM16 5DQ, GB

AGN 28352

OS ESP1996037 EP 0721016 A2 960710

SO Wila-EPZ-1996-H28-T1a

DT Patent

LA Anmeldung in Englisch; Veroeffentlichung in Englisch

DS R DE; R FR; R GB; R IT; R NL

PIT EPA2 EUROPAEISCHE PATENTANMELDUNG

PI EP 721016 A2 960710

OD 960710

AI EP 95-307501 951020

PRAI US 94-327522 941021
US 94-327687 941024
US 95-533582 951018

ABEN Methods for discriminating between fully complementary hybrids and those that differ by one or more base pairs and libraries of unimolecular, double-stranded ***oligonucleotides*** on a solid support. In one embodiment, the present invention provides methods of using nuclease treatment to improve the quality of hybridization ***signals*** on high density ***oligonucleotide*** arrays. In another embodiment, the present invention provides methods of using ligation reactions to improve

the quality of hybridization ***signals*** on high density ***oligonucleotide*** arrays. In yet another embodiment, the present invention provides libraries of unimolecular or intermolecular, double-stranded ***oligonucleotides*** on a solid support. These libraries are useful in pharmaceutical discovery for the screening of numerous biological samples for specific interactions between the double-stranded ***oligonucleotides*** , and peptides, proteins, drugs and RNA. In a related aspect, the present invention provides libraries of conformationally restricted probes on a solid support. The probes are restricted in their movement and flexibility using double-stranded ***oligonucleotides*** as scaffolding. The probes are also useful in various screening procedures associated with drug discovery and diagnosis. The present invention further provides methods for the preparation and screening of the above libraries. <image>

The 5' half of the 20-mer target is complementary to the probes on the chip for which N is a G. The probe:target hybrids for the other three probes have a ***single*** ***base*** ***mismatch*** one base in from the 5' end of the probe. The ligatable 6-mer is complementary to the 3' overhang of. . . The . . . in a 1 ml flow cell containing 10 nM target oligo, 20 nM ligatable 6-mer, and 4000 units of T4 ***DNA*** Ligase (New England Biolabs). The buffer is the buffer recommended by the manufacturer plus 150 mM NaCl. The reaction is. . . after which the chip is vigorously washed with water at 50.degree.C to remove the labelled target molecules. The only fluorescent ***label*** remaining after washing is that of the ligatable 6-mers that have been covalently attached to the probes via the ligation. . . the 5' end (see, supra). For the target used here, G is the complementary base. HYB and LIG are the ***signals*** (fluorescence counts) for the different probes following hybridization and ligation, respectively. HDF and LDF are the discrimination factors (defined as the ***ratio*** of the fluorescence ***signal*** with the perfect match, G, to the ***signal*** with the specified ***mismatch*** base) following hybridization and ligation, respectively. It . . . after hybridization, the extent of target hybridization is very similar for the perfectly complementary probe and the probes containing a ***mismatch*** near the 5' end. The A and C mismatches differ by only 10%, and the maximum difference is only 40%.. . . with the minimum discrimination factor greater than 4. These data indicate that ligation reactions can be performed on covalently attached ***oligonucleotide*** probes on the chip surface, that these reactions are specific for correctly base-paired probe:target hybrids, and that the reaction can be used to improve the discrimination between perfect matches and ***single*** ***base*** mismatches.

In this example, a chip was made with probes having the following ***sequences*** :

P-P-A-A-CGCGCATTCN-5' (denoted CG)

P-P-A-A-ATATAATTCN-5' (denoted AT)

A, T, C, G and N have the same definitions as those set forth in Example I, supra. These probes contain a perfect match and the ***single*** - ***base*** ***mismatch*** ***sequences*** for the following 22-mer target oligos (listed 5' to 3'):

F1-GCGCGTAAGGCCTTCGACGTAG (denoted OH1)

F1-TATATTAAGGCCTTCGACGTAG (denoted OH2)

The 5' end of. . . of OH2 is complementary to the AT probes with N = C. Both OH1 and OH2 have the same 12-mer ***sequence*** at the 3' end. The labelled, ligatable 6-mer used in this example (appropriate for both OH1 and OH2 when hybridized to the CG and AT regions of the chip, respectively) has the following

sequence :

F1-CGAAGG (denoted L6B).

Prior . . . ligation conditions are the same as those used in Example I unless otherwise specified. In particular, 2000 units of T4 ***DNA*** Ligase are used for the reaction here, and the concentration of the ligatable 6-mer is 10 nM rather than 20. . .

The . . . of hybridization to both the CG and AT regions of the chip is analyzed. It is found that the fluorescence ***signal*** in the CG regions (OH1 hybrids) is larger than in the AT regions (OH2 hybrids) by more than a factor of 14. In fact, the perfect match ***signal*** in the CG region is quite strong, but the ***signal*** in the AT region is only slightly greater than twice the background. <table>

Following . . . the chip by combining OH1, OH2, and L6B in 1 ml of ligation buffer and adding 2000 units of T4 ***DNA*** Ligase. The reaction is allowed to proceed for 34 hours at 22.degree.C, and then for another 24 hours at 8.degree.C.. . . It is striking that after the ligation reaction at 8.degree.C, the ***signals*** for OH1 and OH2 differ by only a factor of 1.4, ten times less than the factor of 14 that. . .

In order for the ligation strategy to be useful for unknown or more complex ***DNA*** targets, it is necessary to use a pool of all possible (4096) 6-mers instead of a specific ligatable 6-mer. The. . . are performed using a mixture of A, C, G, and T phosphoramidite, producing a mixture of all possible five base ***sequences*** on each of the four columns. The 6-mers are labelled with fluorescein at the 5' end as the last step. . . 260 nm. The appropriate amounts of each pool is mixed to make a solution that contains all 4096 labelled 6-mer ***oligonucleotides*** .

A chip is made containing 10-mer probes having the following ***sequences***

P-P-C-G-C-G-N.sub1.-N.sub2.-N.sub3.-N.sub4.-N.sub5.-N.sub6.-5'

wherein: N.subi. are A, C, G, or T. In other words, the chip contains 10-mers with all possible (4096). . . the synthesis of the chip, prior to deprotection of the bases. The target oligo is a 22-mer having the following ***sequence*** (listed 5' to 3'):

F1-GCGCGTAAGGCCTTCGACGTAG (OH1)

The chip was initially hybridized with 10 nM OH1 in 6XSSP-T at 22.degree.C for. . . chip is read and analyzed. The only perfect match probe for this target (i.e., PP-CGCGCATTCC-5') has the second highest hybridization ***signal*** . Eight other probes have hybridization ***signals*** that are within a factor of 4 of the perfect match ***signal*** . The other three

probes with a ***single*** ***base*** ***mismatch*** at the 5' end have discrimination factors of 2.0, 2.6, and 3.5, for G, A, and T, respectively. Other ***single*** ***base*** mismatches at positions in from the 5' end of the probe give ***signals*** that are considerably smaller. The chip is washed with water to remove the hybridized target.

The . . . the ligation buffer for 11 hours at 22.degree.C (no ligase at this stage). The perfect match probe gives the highest ***signal*** by a factor of 2.4. Five probes have ***signals*** within a factor of 4 of the perfect match ***signal*** . The other three probes with a ***single*** ***base*** ***mismatch*** at the 5' end have discrimination factors of 3.0, 3,6, and 8.0, for G, A, and T, respectively.

The ligation reaction is initiated by the addition of 2000 units of T4 ***DNA*** ligase to the solution containing OH1 and the pool of 6-mers. The reaction is allowed to proceed for 23 hours. . . chip with water at about 45.degree.C for five minutes, the chip is read. After ligation, no other probes have hybridization ***signals*** that are within a factor of 4 of the perfect match ***signal*** . The three 5' ***single*** ***base*** ***mismatch*** probes all have discrimination factors greater than 12. Thus, with a complex chip containing 4096 probes with all possible 6-mer ***sequences*** at the 5' end, and using a pool of all possible ligatable 6-mers, the ligation reaction is still specific for the perfectly complementary probe and affords considerable increases in the discrimination between perfect matches and ***single*** - ***base*** mismatches.

In . . . example, a chip was made using the tiling strategy (A, C, G, T -containing probes for each base in the ***sequence***) described above that covers a 50 base region of the protease ***gene*** of HIV-1 (SF2 strain). The probes are 11-mers, linked to the glass support by three PEG linkers. The substitution position. . . and ligation, the chip is phosphorylated using T4 polynucleotide kinase for 5 hours at 37.degree.C. The target is a 75-mer ***oligonucleotide*** (denoted Hpro1), labelled at the 5' end with fluorescein, that spans the complementary 50 base region on the chip.

The . . . then carried out with 10 nM Hpro1, 1.6 .mu.M 6-mer pool (0.4 nM per oligo), and 2000 units of T4 ***DNA*** Ligase in 1 ml of ligation buffer. The ligation reaction is allowed to proceed for 25 hours at 8.degree.C, then. . . and finally 4 days at 8.degree.C. At intervals of 1 to 2 days, the solution is supplemented with additional T4 ***DNA*** Ligase. Following the ligation reaction, the chip is washed vigorously with water at about 45.degree.C for 10 minutes, leaving only. . .

The . . . hybridization and ligation reactions are analyzed in terms of the ability to make a correct base call from the

fluorescence ***signal*** measured on the chip. In particular, the ***signal*** is compared between the four probes that differ by a ***single*** ***base*** at a given position within the 11-mer, with the rest of the 11-mer being perfectly complementary to a specific region of the target ***sequence*** . For the purposes of this experiment, a base identification is said to be made if the ***signal*** in at least one of the four probe regions is greater than the ***signal*** in a nearby region that has no ***oligonucleotide*** probes (the background) by at least 5 counts (the background counts are usually about 2 - 6 counts), and if the ***signal*** in one of the four regions is greater than that in the other three related regions by at least a factor of 1.2. If none of the four ***signals*** are larger than the other three by a factor of at least 1.2, a multiple base ambiguity results. If the most intense hybridization ***signal*** (by a factor of at least 1.2) is for a probe that is not perfectly complementary to the target ***sequence*** , then a miscall results.

Following . . . or miscalls. These results indicate that the ligation reaction with the full pool of 6-mers can be used to specifically ***label*** hybrids between relatively complex targets and ***arrays*** of ***oligonucleotide*** probes. It is interesting to note that the pattern of ligation (stronger or weaker ***signals*** , better or worse discrimination) is not in general the same as the pattern of hybridization. This suggests that these two approaches may be used as complementary tools to obtain ***sequence*** information with ***arrays*** of ***oligonucleotide*** probes. For example, probes that produce large hybridization ***signals*** , but are poorly discriminated may be better treated using a ligation step. And probes that do not hybridize well to a particular complementary target (leading to a ***signal*** that is too small relative to the background) may ligate well enough to be clearly detected (as also suggested by. . .

C. PREPARATION OF UNIMOLECULAR, DOUBLE-STRANDED ***OLIGONUCLEOTIDES***

This example illustrates the general synthesis of an ***array*** of unimolecular, double-stranded ***oligonucleotides*** on a solid support.

Unimolecular double stranded ***DNA*** molecules were synthesized on a solid support using standard light-directed methods (VLSIPS.trade. protocols). Two hexaethylene glycol (PEG) linkers were used to covalently attach the synthesized ***oligonucleotides*** to the derivatized glass surface. Synthesis of the first (inner) strand proceeded one nucleotide at a time using repeated cycles. . . Upon completion of the inner strand, another MeNPoc-protected PEG linker was covalently

attached to the 5' end of the surface-bound ***oligonucleotide*** . After addition of the internal PEG linker, the PEG is photodeprotected, and the synthesis of the second strand proceeded in the normal fashion. Following the synthesis cycles, the ***DNA*** bases were deprotected using standard protocols. The ***sequence*** of the second (outer) strand, being complementary to that of the inner strand, provided molecules with short, hydrogen bonded, unimolecular. . .

An ***array*** of 16 different molecules were synthesized on a derivatized glass slide in order to determine whether short, unimolecular ***DNA*** structures could be formed on a surface and whether they could adopt structures that are recognized by proteins. Each of. . . S is the solid surface having silyl groups, P is a PEG linker, A, C, G, and T are the ***DNA*** nucleotides, and F is a fluorescent tag. The ***DNA*** ***sequence*** is listed from the 3' to the 5' end (the 3' end of the ***DNA*** molecule is attached to the solid surface via a silyl group and 2 PEG linkers). The sixteen molecules synthesized on. . .

This example illustrates the ability of a library of surface-bound, unimolecular, double-stranded ***oligonucleotides*** to exist in duplex form and to be recognized and bound by a protein.

A . . . complementary to the inner strand (these molecules will be referred to as DS, double-stranded, below). One of the four DS ***oligonucleotides*** has a ***sequence*** that is recognized by the restriction enzyme EcoR1. If the molecule can loop back and form a ***DNA*** duplex, it should be recognized and cut by the restriction enzyme, thereby releasing the fluorescent tag. Thus, the action of the enzyme provided a functional test for ***DNA*** structure, and also served to demonstrate that these structures can be recognized at the surface by proteins. The remaining 12. . . (referred to as SS, single-stranded, below). Of these, three had an outer strand and three had an inner strand whose ***sequence*** was an EcoR1 half-site (the ***sequence*** on one strand was correct for the enzyme, but the other half was not). The solid support with an ***array*** of molecules on the surface is referred to as a "chip" for the purposes of the following discussion. The presence. . . fluorescence from the molecules on the chip surface (e.g. "reading" the chip) upon treatment with enzymes that can cut the ***DNA*** and release the fluorescent tag at the 5' end.

The . . . different enzymes used to characterize the structure of the molecules on the chip were:

1) Mung Bean Nuclease - ***sequence*** independent, single-strand specific ***DNA*** endonuclease;

2) DNase I - ***sequence*** independent, double-strand

specific endonuclease;

3) EcoR1 - restriction endonuclease that recognizes the ***sequence*** (5'-3')

GAATTC in double stranded ***DNA*** , and cuts between the G and the first A. Mung Bean Nuclease and EcoR1 were obtained from New England Biolabs,. . .

Upon treatment of the chip with the enzyme EcoR1, the fluorescence ***signal*** in the DS EcoR1 region and the 3 SS regions with the EcoR1 half-site on the outer strand was reduced. . . least 5 times greater than for the other regions of the chip, indicating that the action of the enzyme is ***sequence*** specific on the chip. It was not possible to determine if the factor is greater than 5 in these preliminary. . .

The reduction in ***signal*** in the 3 SS regions with the EcoR1 half-site on the outer strand indicated either that the enzyme cuts single-stranded ***DNA*** with a particular ***sequence*** , or that these molecules formed a double-stranded structure that was recognized by the enzyme. The molecules on the chip surface. . . 3 SS regions with the EcoR1 half-site on the outer strand, such a bimolecular double-stranded region would have the correct ***sequence*** and structure to be recognized by EcoR1. However, it would differ from the unimolecular double-stranded molecules in that the inner. . . by a single-strand specific endonuclease such as Mung Bean Nuclease. Therefore, it was possible to distinguish unimolecular from bimolecular double-stranded ***DNA*** molecules on the surface by their ability to be cut by single and double-strand specific endonucleases.

In . . . to identify unimolecular double-stranded molecules, the chip was first exhaustively treated with Mung Bean Nuclease. The reduction in the fluorescence ***signal*** was greater by about a factor of 2 for the SS regions of the chip, including those with the EcoR1. . . I (which cuts all remaining double-stranded molecules) or EcoR1 (which should cut only the remaining double-stranded molecules with the correct ***sequence***). Upon treatment with DNase I, the fluorescence ***signal*** in the 4 DS regions was reduced by at least 5-fold more than the ***signal*** in the SS regions. Upon EcoR1 treatment, the ***signal*** in the single DS region with the correct EcoR1 ***sequence*** was reduced by at least a factor of 3 more than the ***signal*** in any other region on the chip. Taken together, these results indicated that the surface-bound molecules synthesized with two complementary. . . structures that were resistant to a single-strand specific endonuclease and were recognized by both a double-strand specific endonuclease, and a ***sequence*** -specific restriction

enzyme.

A glass coverslip having aminopropylsilane spacer groups can be further derivatized on the amino groups with a poly-A ***oligonucleotide*** comprising nine adenosine monomers using VLSIPS.trade. ("light-directed") methods. The tenth adenine monomer to be added will be a 5'-aminopropyl-functionalized phosphoramidite. . .

The present invention provides greatly improved methods and apparatus for the study of nucleotide ***sequences*** and ***nucleic*** ***acid*** interactions with other molecules. It is to be understood that the above description is intended to be illustrative and not. . .

CLMEN 1. A method for ***sequencing*** a target ***nucleic*** ***acid*** , said method comprising:

(a) combining:

(i) a substrate comprising an ***array*** of chemically synthesized and positionally distinguishable ***oligonucleotides*** each of which is complementary to a defined subsequence of preselected length; and

(ii) a target ***nucleic*** ***acid*** ; thereby forming target- ***oligonucleotide*** hybrid complexes of complementary subsequences of known ***sequence*** ;

. A ***computer*** 100 is used to design ***arrays*** of biological polymers such as RNA or ***DNA*** . The ***computer*** 100 may be, for example, an appropriately programmed Sun Workstation or personal ***computer*** or workstation, such as an IBM PC equivalent, including appropriate memory and a CPU as shown in Figs. 1 and 2. The ***computer*** system 100 obtains inputs from a user regarding characteristics of a ***gene*** of interest, and other inputs regarding the desired features of the ***array*** . Optionally, the ***computer*** system may obtain information regarding a specific genetic ***sequence*** of interest from an external or internal database 102 such as GenBank. The output of the ***computer*** system 100 is a set of chip design ***computer*** files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated ***computer*** files.

The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of ***arrays*** of molecules such as ***DNA*** . The system or process 106 may include the hardware necessary to manufacture masks 110 and also the necessary ***computer*** hardware and ***software*** 108 necessary to lay the mask patterns out on the mask in an efficient manner. As with the other features. . . system 106 generates masks 110 or other synthesis patterns such as chrome-on-glass masks for use in the fabrication of polymer ***arrays*** .

The . . . the chips from system 100, are used in a synthesis system 112. Synthesis system 112 includes the necessary hardware and ***software*** used to fabricate ***arrays*** of polymers on a substrate or chip 114. For example, synthesizer 112 includes a light source 116 and a chemical. . . to deprotected regions, as well as for washing and other operations. All operations are preferably directed by an appropriately programmed ***computer*** 119, which may or may not be the same ***computer*** as the ***computer*** (s) used in mask design and mask making.

The . . . may not be complementary to one or more of the molecules on the substrate. The receptors are marked with a ***label*** such as a fluorescein ***label*** (indicated by an asterisk in Fig. 3) and placed in scanning system 120. Scanning system 120 again operates under the direction of an appropriately programmed digital ***computer*** 122, which also may or may not be the same ***computer*** as the ***computers*** used in synthesis, mask making, and mask design. The scanner 120 includes a detection device 124 such as a confocal. . . The output of scanner 120 is an image file(s) 124 indicating, in the case of fluorescein labeled receptor, the fluorescence

intensity (photon counts or other related measurements, such as voltage) as a function of position on the substrate. Since higher photon counts will be observed where the labeled receptor has bound more strongly to the ***array*** of polymers, and since the monomer ***sequence*** of the polymers on the substrate is known as a function of position, it becomes possible to determine the ***sequence*** (s) of polymer(s) on the substrate that are complementary to the receptor.

The . . . and analysis methods of the present invention. Again, the analysis system may be any one of a wide variety of ***computer*** system(s), but in a preferred embodiment the analysis system is based on a Sun Workstation or equivalent. The present invention. . . providing appropriate output 128. The present invention may further be used to identify specific mutations in a receptor such as ***DNA*** or RNA.

Fig. 4 provides a simplified illustration of the overall ***software*** system used in the operation of one embodiment of the invention. As shown in Fig. 4, in some cases (such as ***sequence*** checking systems) the system first identifies the genetic ***sequence*** (s) or targets that would be of interest in a particular analysis at step 202. The ***sequences*** of interest may, for example, be normal or mutant portions of a ***gene*** , ***genes*** that identify heredity, or provide forensic information, or be all possible n-mers (where n represents the length of the ***nucleic*** ***acid***). ***Sequence*** selection may be provided via manual input of text files or may be from external sources such as GenBank. At step 204 the system evaluates the ***gene*** to determine or assist the user in determining which probes would be desirable on the chip, and provides an appropriate "layout" on the chip for the probes. The chip usually includes probes that are complementary to a reference ***nucleic*** ***acid*** ***sequence*** which has a known ***sequence*** . A wild-type probe is a probe that will ideally hybridize with the reference ***sequence*** and thus a wild-type ***gene*** (also called the chip wild-type) would ideally hybridize with wild-type probes on the chip. The target ***sequence*** is substantially similar to the reference ***sequence*** except for the presence of mutations, insertions, deletions, and the like. The layout implements desired characteristics such as arrangement on the chip that permits "reading" of genetic ***sequence*** and/or minimization of edge effects, ease of synthesis, and the like.

Fig. 5 illustrates the global layout of a chip in a particular embodiment used for ***sequence*** checking applications. Chip 114 is composed of multiple units where each unit may contain different tilings for the chip wild-type ***sequence*** . Unit 1 is shown in greater detail and shows that each unit is composed

of multiple cells which are areas. . . a "blank" cell 224. Cell 220 contains a wild-type probe that is the complement of a portion of the wild-type ***sequence*** . Cells 222 contain "mutation" probes for the wild-type ***sequence*** . For example, if the wild-type probe is 3'-ACGT, the probes 3'-ACAT, 3'-ACCT, 3'-ACGT, and 3'-ACTT may be the "mutation" probes.. . . the chip in this area. Thus, the blank cell provides an area that can be used to measure the background ***intensity*** .

In one embodiment, numerous tiling processes are available including ***sequence*** tiling, block tiling, and opt-tiling, as described below. Of course a wide range of layout strategies may be used according. . . of the invention. For example, the probes may be tiled on a substrate in an apparently random fashion where a ***computer*** system is utilized to keep track of the probe locations and correlate the data obtained from the substrate.

Opt-tiling is the process of tiling additional probes for suspected mutations. As a simple example of opt-tiling, suppose the wild-type target ***sequence*** is 5'-ACGTATGCA-3' and it is suspected that a mutant ***sequence*** has a possible T base mutation at the underlined base position. Suppose further that the chip will be synthesized with. . .

In . . . row of the probes (along with one probe below each of the four wild-type probes) should bind to the target ***DNA*** ***sequence*** . However, if the target ***sequence*** has a T base mutation as suspected, the labeled mutant ***sequence*** will not bind that strongly to the probes in the columns around column 3. For example, the mutant receptor that. . .

Again referring to Fig. 4, at step 206 the masks for the synthesis are designed. At step 208 the ***software*** utilizes the mask design and layout information to make the ***DNA*** or other polymer chips. This ***software*** 208 will control relative translation of a substrate and the mask, the flow of desired reagents through a flow cell, the synthesis temperature of the flow cell, and other parameters. At step 210, another piece of ***software*** is used in scanning a chip thus synthesized and exposed to a labeled receptor. The ***software*** controls the scanning of the chip, and stores the data thus obtained in a file that may later be utilized to extract ***sequence*** information.

At step 212 a ***computer*** system according to the present invention utilizes the layout information and the fluorescence information to evaluate the hybridized ***nucleic*** ***acid*** probes on the chip. Among the important pieces of information obtained from probe ***arrays*** are the identification of mutant receptors and determination of genetic ***sequence*** of a particular receptor.

Fig. 6 illustrates the binding of a particular target ***DNA*** to an ***array*** of ***DNA*** probes 114. As shown in this simple example, the following probes are formed in the ***array*** (only one probe is shown for the wild-type probe): <image> As shown, the set of probes differ by only one base so the probes are designed to determine the identity of the base at that position in the ***nucleic*** ***acid*** ***sequence*** .

When a fluorescein-labeled (or otherwise marked) target with the ***sequence*** 5'-TCTTGCA is exposed to the ***array*** , it is complementary only to the probe 3'-AGAACGT, and fluorescein will be primarily found on the surface of the chip. . . located. Thus, for each set of probes that differ by only one base, the image file will contain four fluorescence ***intensities*** , one for each probe. Each fluorescence ***intensity*** can therefore be associated with the base of each probe that is different from the other probes. Additionally, the image file will contain a "blank" cell which can be used as the fluorescence ***intensity*** of the background. By analyzing the five fluorescence ***intensities*** associated with a specific base location, it becomes possible to extract ***sequence*** information from such ***arrays*** using the methods of the invention disclosed herein.

Fig. 7 illustrates probes arranged in lanes on a chip. A reference ***sequence*** is shown with five interrogation positions marked with number subscripts. An interrogation position is a base position in the reference ***sequence*** where the target ***sequence*** may contain a mutation or otherwise differ from the reference ***sequence*** . The chip may contain five probe cells that correspond to each interrogation position. Each probe cell contains a set of. . . probes that have a common base at the interrogation position. For example, at the first interrogation position, I.sub1., the reference ***sequence*** has a base T. The wild-type probe for this interrogation position is 3'-TGAC where the base A in the probe is complementary to the base at the interrogation position in the reference ***sequence*** .

Similarly, . . . I.sub1.. The four mutant probes are 3'-TGAC, 3'-TGCC, 3'-TGGC, and 3'-TGTC. Each of the four mutant probes vary by a ***single*** ***base*** at the interrogation position. As shown, the wild-type and mutant probes are arranged in lanes on the chip. One of. . .

Fig. 8 illustrates a hybridization pattern of a target on a chip with a reference ***sequence*** as in Fig. 7. The reference ***sequence*** is shown along the top of the chip for ***comparison*** . The chip includes a WT-lane (wild-type), an A-lane, a C-lane, a G-lane, and a T-lane (or U). Each lane is a

row of cells containing probes. The cells in the WT-lane contain probes that are complementary to the reference ***sequence*** . The cells in the A-, C-, G-, and T-lanes contain probes that are complementary to the reference ***sequence*** except that the named base is at the interrogation position.

In one embodiment, the hybridization of probes in a cell is determined by the fluorescent ***intensity*** (e.g., photon counts) of the cell resulting from the binding of marked target ***sequences*** . The fluorescent ***intensity*** may vary greatly among cells. For simplicity, Fig. 8 shows a high degree of hybridization by a cell containing a. . .

In practice, the fluorescent ***intensities*** of cells near an interrogation position having a mutation are relatively dark creating "dark regions" around a mutation. The lower fluorescent ***intensities*** result because the cells at interrogation positions near a mutation do not contain probes that are perfectly complementary to the target ***sequence*** ; thus, the hybridization of these probes with the target ***sequence*** is lower. For example, the relative ***intensity*** of the cells at interrogation positions I.sub3. and I.sub5. may be relatively low because none of the probes therein are complementary to the target ***sequence*** .

II. ***Intensity*** ***Ratio*** Method

The ***intensity*** ***ratio*** method is a method of calling bases in a sample ***nucleic*** ***acid*** ***sequence*** . The ***intensity*** ***ratio*** method is most accurate when there is good discrimination between the fluorescence ***intensities*** of hybrid matches and hybrid mismatches. If there is insufficient discrimination, the ***intensity*** ***ratio*** method assigns a corresponding ambiguity code to the unknown base.

For simplicity, the ***intensity*** ***ratio*** method will be described as being used to identify one unknown base in a sample ***nucleic*** ***acid*** ***sequence*** . In practice, the method is used to identify many or all the bases in a ***nucleic*** ***acid*** ***sequence*** .

The . . . location where a labeled receptor should not bind to the chip since no probe is present. For example, suppose a ***DNA*** ***sequence*** of interest or target ***sequence*** contains the ***sequence*** 5'-AGAACCTGC-3' with a possible mutation at the underlined base position. Suppose that 5-mer probes are to be synthesized for the target ***sequence*** . A representative wild-type probe of 5'-TTGGA is complementary to the region of the ***sequence*** around the possible mutation. The "mutation" probes will be the same as the wild-type probe except for a different base. . .

If the fluorescently marked sample ***sequence*** is exposed

to the above four mutation probes, the ***intensity*** should be highest for the probe that binds most strongly to the sample ***sequence*** . Therefore, if the probe 3'-TTTGA shows the highest ***intensity*** , the unknown base in the sample will generally be called an A mutation because the probes are complementary to the sample ***sequence*** .

The identity of the unknown base is preferably determined by evaluating the relative fluorescence ***intensities*** of up to four of the mutation probes, and the "blank" cell. Because each mutation probe is identifiable by the mutation base, a mutation probe's ***intensity*** will be referred to as the "base ***intensity*** " of the mutation base.

As a simple example of the ***intensity*** ***ratio*** method, suppose a ***gene*** of interest (target) is an HIV protease ***gene*** with the ***sequence*** 5'-ATGTGGACAGTTGTA-3' (SEQ ID NO:1). Suppose further that a sample ***sequence*** is suspected to have the same ***sequence*** as the target ***sequence*** except for a mutation of base C to base T at the underlined base position. Although hundreds of probes may be synthesized on the chip, the complementary mutation probes synthesized to detect a mutation in the sample ***sequence*** at the suspected mutation position may be as follows:

3'-TATC
3'-TCTC
3'-TGTC (wild-type)
3'-TTTC

The mutation probe 3'-TGTC is also the wild-type probe as it should bind most strongly with the target ***sequence*** . After the sample ***sequence*** is labeled, hybridized on the chip, and scanned, suppose the following fluorescence ***intensities*** were obtained:

3'-TATC -> 45
3'-TCTC -> 8
3'-TGTC -> 32
3'-TTTC -> 12

where the ***intensity*** is measured by the photon count detected by the scanner. The "blank" cell had a fluorescence ***intensity*** of 2. The photon counts in the examples herein are representative (not actual data) and provided for illustration purposes. In. . .

Although each fluorescence ***intensity*** is from a probe, the probes may be characterized by their unique mutation base so the bases may be said to have the following ***intensities*** :

A -> 45
C -> 8
G -> 32

T -> 12

Thus, base A will be described as having an ***intensity*** of 45, which corresponds to the ***intensity*** of the mutation probe with the mutation base A.

Initially, each mutation base ***intensity*** is reduced by the background or "blank" cell ***intensity*** . This is done as follows: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> Then, the base ***intensities*** are sorted in descending order of ***intensity*** . The above bases would be sorted as follows:

A -> 43
G -> 30
T -> 10
C -> 6

Next, the highest ***intensity*** base is compared to the second highest ***intensity*** base. Thus, the ***ratio*** of the ***intensity*** of base A to the ***intensity*** of base G is calculated as follows: A:G = 43 / 30 = 1.4. The ***ratio*** A:G is then compared to a predetermined ***ratio*** cutoff, which is a number that specifies the ***ratio*** required to identify the unknown base. For example, if the ***ratio*** cutoff is 1.2, the ***ratio*** A:G is greater than the ***ratio*** cutoff (1.4 > 1.2) and the unknown base is called by the mutation probe containing the mutation A. As probes are complementary to the sample ***sequence*** , the sample ***sequence*** is called as having a mutation T, resulting in a called sample ***sequence*** of 5'-ATGTGGATAGTTGTA-3' (SEQ ID NO:2).

As another example, suppose everything else is the same as in the previous example except that the sorted background adjusted ***intensities*** were as follows:

C -> 42
A -> 40
G -> 10
T -> 8

The ***ratio*** of the highest ***intensity*** base to the second highest ***intensity*** base (C:A) is 1.05. Because this ***ratio*** is not greater than the ***ratio*** cutoff of 1.2, the unknown base will be called as being ambiguously one of two or more bases as follows.. . .

The second highest ***intensity*** base is then compared to the third highest base. The ***ratio*** of A:G is 4. The ***ratio*** of A:G is then compared to the ***ratio*** cutoff of 1.2. As the ***ratio*** A:G is greater than the ***ratio*** cutoff (4 > 1.2), the unknown base is called by the mutation probes containing the mutations C or A. As probes are

complementary to the sample ***sequence*** , the sample ***sequence*** is called as having either a mutation G or T, resulting in a sample ***sequence*** of 5'-ATGTGGAKAGTTGTA-3' (SEQ ID NO:3) where K is the IUPAC code for G or T(U).

The ***ratio*** cutoff in the previous examples was equal to 1.2. However, the ***ratio*** cutoff will generally need to be adjusted to produce optimal results for the specific chip design and wild-type target. Also, although the ***ratio*** cutoff used has been the same for each ***ratio*** ***comparison*** , the ***ratio*** cutoff may vary depending on whether the ***ratio*** ***comparisons*** involve the highest, second highest, third highest, etc. ***intensity*** base.

Fig. 9 illustrates the high level flow of the ***intensity*** ***ratio*** method. At step 302 the four base ***intensities*** are adjusted by subtracting the background or "blank" cell ***intensity*** from each base ***intensity*** . Preferably, if a base ***intensity*** is then less than or equal to zero, the base ***intensity*** is set equal to a small positive number to prevent division by zero or negative numbers in future calculations.

At step 304 the base ***intensities*** are sorted by ***intensity*** . Each base is then associated with a number from 1 to 4. The base with the highest ***intensity*** is 1, second highest 2, third highest 3, and fourth highest 4. Thus, the ***intensity*** of base 1 >= base 2 >= base 3 >= base 4.

At step 306 the highest ***intensity*** base (base 1) is checked to see if it has sufficient ***intensity*** to call the unknown base. The ***intensity*** is checked by determining if the ***intensity*** of base 1 is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the ***intensity*** a base ***intensity*** must be over the background ***intensity*** in order to correctly call the unknown base. Thus, the background adjusted base ***intensity*** must be greater than the background difference cutoff or the unknown is not callable.

If the ***intensity*** of base 1 is not greater than the background difference cutoff, the unknown base is assigned the code N (insufficient ***intensity***) as shown at step 308. Otherwise, the ***ratio*** of the ***intensity*** of base 1 to base 2 is calculated as shown at step 310.

At step 312 the ***ratio*** of ***intensity*** of bases 1:2 is compared to the ***ratio*** cutoff. If the ***ratio*** 1:2 is greater than the ***ratio*** cutoff, the unknown base is called as the complement of the highest ***intensity*** base (base 1) as shown at step 314. Otherwise, the ***ratio*** of the ***intensity*** of base 2 to base 3

is calculated as shown at step 316.
At step 318 the ***ratio*** of ***intensity*** of bases 2:3 is compared to the ***ratio*** cutoff. If the ***ratio*** 2:3 is greater than the ***ratio*** cutoff, the unknown base is called as being an ambiguity code specifying the complements of the highest or second highest ***intensity*** bases (base 1 or 2) as shown at step 320. Otherwise, the ***ratio*** of the ***intensity*** of base 3 to base 4 is calculated as shown at step 322.
At step 324 the ***ratio*** of ***intensity*** of bases 3:4 is compared to the ***ratio*** cutoff. If the ***ratio*** 3:4 is greater than the ***ratio*** cutoff, the unknown base is called as being an ambiguity code specifying the complements of the highest, second highest, or. . .
The advantage of the ***intensity*** ***ratio*** method is that it is very accurate when there is good discrimination between the fluorescence ***intensities*** of hybrid matches and hybrid mismatches. However, if the base corresponding to a correct hybrid gives a lower ***intensity*** than a ***mismatch*** (e.g., as a result of cross-hybridization), incorrect identification of the base will result. For this reason, however, the method is useful for comparative assessment of hybridization quality and as an indicator of ***sequence*** -specific problem spots. For example, the ***intensity*** ***ratio*** method has been used to determine that ambiguities and miscalls tend to be very different from ***sequence*** to ***sequence*** , and reflect predominantly the composition and repetitiveness of the ***sequence*** . It has also been used to assess improvements obtained by varying hybridization conditions, sample preparation, and post-hybridization treatments (e.g., RNase. . .
The reference method is a method of calling bases in a sample ***nucleic*** ***acid*** ***sequence*** . The reference method depends very little on discrimination between the fluorescence ***intensities*** of hybrid matches and hybrid mismatches, and therefore is much less sensitive to cross-hybridization. The method compares the probe ***intensities*** of a reference ***sequence*** to the probe ***intensities*** of a sample ***sequence*** . Any significant changes are flagged as possible mutations. There are two implementations of the reference method disclosed herein. For simplicity, the reference method will be described as being used to identify one unknown base in a sample ***nucleic*** ***acid*** ***sequence*** . In practice, the method is used to identify many or all the bases in a ***nucleic*** ***acid*** ***sequence*** .
The unknown base will be called by comparing the probe ***intensities*** of a reference ***sequence*** to the probe

intensities of a sample ***sequence*** . Preferably, the probe ***intensities*** of the reference ***sequence*** and the sample ***sequence*** are from chips having the same chip wild-type. However, the reference ***sequence*** may or may not be exactly the same as the chip wild-type, as it may have mutations.

The bases at the same position in the reference and sample ***sequences*** will each be associated with up to four mutation probes and a "blank" cell. The unknown base in the sample ***sequence*** is called by comparing probe ***intensities*** of the sample ***sequence*** to probe ***intensities*** of the reference ***sequence*** . For example, suppose the chip wild-type contains the ***sequence*** 5'-AGACCTTGC-3' and it is suspected that the sample has a possible mutation at the underlined base position, which is the unknown base that will be called by the reference method. The "mutation" probes for the sample ***sequence*** may be as follows: 3'-GAAA, 3'-GCAA, 3'-GGAA, and 3'-GTAA, where 3'-GGAA is the wild-type probe. Suppose further that a reference ***sequence*** , which differs from the chip wild-type by one base mutation, has the ***sequence*** 5'-AGACATTGC-3' where the mutation base is underlined. The "mutation" probes for the reference ***sequence*** may be as follows: 3'-TGAAA, 3'-TGCAA, 3'-TGGAA, and 3'-TGTAA, where 3'-TGTAA is the reference wild-type probe since the reference ***sequence*** is known. Although generally the sample and reference ***sequences*** were tiled with the same chip wild-type, this is not required, and the tiling methods do not have to be. . . probe lengths in the example. Thus, the unknown base will be called by comparing the "mutation" probes of the sample ***sequence*** to the "mutation" probes of the reference ***sequence*** . As before, because each mutation probe is identifiable by the mutation base, the mutation probes' ***intensities*** will be referred to as the "base ***intensities*** " of their respective mutation bases.

As a simple example of one implementation of the reference method, suppose a ***gene*** of interest (target) has the ***sequence*** 5'-AAAACTGAAAA-3' (SEQ ID NO:4). Suppose a reference ***sequence*** has the ***sequence*** 5'-AAAACCGAAAA-3' (SEQ ID NO:5), which differs from the target ***sequence*** by the underlined base. The reference ***sequence*** is marked and exposed to probes on a chip with the target ***sequence*** being the chip wild-type. Suppose further that a sample ***sequence*** is suspected to have the same ***sequence*** as the target ***sequence*** except for a mutation at the underlined base position in 5'-AAAACTGAAAA-3' (SEQ ID NO:4). The sample ***sequence*** is also marked and exposed to probes on a chip with the target

sequence being the chip wild-type. After hybridization and scanning, the following probe ***intensities*** (not actual data) were found for the respective complementary probes: <table> Although each fluorescence ***intensity*** is from a probe, the probes may be identified by their unique mutation base so the bases may be said to have the following ***intensities*** : <table> Thus, base A of the reference ***sequence*** will be described as having an ***intensity*** of 12, which corresponds to the ***intensity*** of the mutation probe with the mutation base A. The reference method will now be described as calling the unknown base in the sample ***sequence*** by using these ***intensities*** .

At step 402 the four base ***intensities*** of the reference and sample ***sequences*** are adjusted by subtracting the background or "blank" cell ***intensity*** from each base ***intensity*** . Each set of "mutation" probes has an associated "blank" cell. Suppose that the reference "blank" cell ***intensity*** is 1 and the sample "blank" cell ***intensity*** is 2. The base ***intensities*** are then background subtracted as follows: <table> Preferably, if a base ***intensity*** is then less than or equal to zero, the base ***intensity*** is set equal to a small positive number to prevent division by zero or negative numbers in future calculations.

For identification, the position of each base of interest in the reference and sample ***sequences*** is placed in column 1 of the analysis table. Also, since the reference ***sequence*** is a known ***sequence*** , the base at this position is known and is referred to as the reference wild-type. The reference wild-type is placed. . .

At step 404 the base ***intensity*** associated with the reference wild-type (column 2 of the analysis table) is checked to see if it has sufficient ***intensity*** to call the unknown base. In this example, the reference wild-type is C. However, the base ***intensity*** associated with the wild-type is the G base ***intensity*** , which is 79 in this example. This is because the base ***intensities*** actually represent the complementary "mutation" probes. The G base ***intensity*** is checked by determining if its ***intensity*** is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the ***intensity*** the base ***intensities*** must be above the background ***intensity*** in order to correctly call the unknown base. Thus, the base ***intensity*** associated with the reference wild-type must be greater than the background difference cutoff or the unknown base is not callable.. . .

If the background difference cutoff is 5, the base ***intensity*** associated with the reference wild-type has sufficient ***intensity*** (79 > 5) so a P (pass) is placed in column 3 of the analysis table as shown at step. . .
At step 408 the ***ratio*** of the base ***intensity*** associated with the reference wild-type to each of the possible bases are calculated. The ***ratio*** of the base ***intensity*** associated with the reference wild-type to itself will be 1 and the other ***ratios*** will usually be greater than 1. The base ***intensity*** associated with the reference wild-type is G so the following ***ratios*** are calculated: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> These ***ratios*** are placed in columns 4 through 7 of the analysis table, respectively.
At step 410 the highest base ***intensity*** associated with the sample ***sequence*** is checked to see if it has sufficient ***intensity*** to call the unknown base. The highest base ***intensity*** is checked by determining if the ***intensity*** is greater than the background difference cutoff. Thus, the highest base ***intensity*** must be greater than the background difference cutoff or the unknown base is not callable.
Again, if the background difference cutoff is 5, the highest base ***intensity*** , which is G in this example, has sufficient ***intensity*** (58 > 5) so a P (pass) is placed in column 8 of the analysis table as shown at step. . .
At step 414 the ***ratios*** of the highest base ***intensity*** of the sample to each of the possible bases are calculated. The ***ratio*** of the highest base ***intensity*** to itself will be 1 and the other ***ratios*** will usually be greater than 1. Thus, the highest base ***intensity*** is G so the following ***ratios*** are calculated: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> These ***ratios*** are placed in columns 9 through 12 of the analysis table, respectively.
At step 416 if both the reference and sample ***sequence*** probes failed to have sufficient ***intensity*** to call the unknown base, meaning there is an 'F' in columns 3 and 8 of the analysis table, the unknown base is assigned the code N (insufficient ***intensity***) as shown at step 418. An 'N' is placed in column 17 of the analysis table. Additionally, a confidence code. . .
At step 420 if only the reference ***sequence*** probes failed to have sufficient ***intensity*** to call the unknown base, meaning there is an 'F' in column 3 and a 'P' in column 8 of the

analysis table, the unknown base is assigned the code N (insufficient ***intensity***) as shown at step 422. An 'N' is placed in column 17 and a confidence code of 4 is placed. . .
At step 424 if only the sample ***sequence*** probes failed to have sufficient ***intensity*** to call the unknown base, meaning there is a 'P' in column 3 and a 'F' in column 8 of the analysis table, the unknown base is assigned the code N (insufficient ***intensity***) as shown at step 426. An 'N' is placed in column 17 and a confidence code of 2 is placed. . .
In this example, both the reference and sample ***sequence*** probes have sufficient ***intensity*** to call the unknown base. At step 428 the ***ratios*** of the reference ***ratios*** to the sample ***ratios*** for each base type are calculated. Thus, the ***ratio*** A:A (column 4 to column 9) is placed in column 13 of the analysis table. The ***ratio*** C:C (column 5 to column 10) is placed in column 14 of the analysis table. The ***ratio*** G:G (column 6 to column 11) is placed in column 15 of the analysis table. Lastly, the ***ratio*** T:T (column 7 to column 12) is placed in column 16 of the analysis table. These ***ratios*** are calculated as follows: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> The unknown base is called by comparing these ***ratios*** of ***ratios*** to two predetermined values as follows.
At step 430 if all the ***ratios*** of ***ratios*** (columns 13 to 16 of the analysis table) are less than a predetermined lower ***ratio*** cutoff, the unknown base is assigned the code of the reference wild-type as shown at step 432. Thus, the code. . .
At step 434 if all the ***ratios*** of ***ratios*** are less than a predetermined upper ***ratio*** cutoff, the unknown base is assigned an ambiguity code that indicates the unknown base may be any one of the bases that has a complementary ***ratio*** of ***ratios*** greater than the lower ***ratio*** cutoff and less than the upper ***ratio*** cutoff as shown at step 436. Thus, if the ***ratio*** of ***ratios*** for A:A, C:C and G:G are all greater than the lower ***ratio*** cutoff and less than the upper ***ratio*** cutoff, the unknown base would be assigned the code B (meaning "not A"). This is because the ***ratios*** of ***ratios*** are complementary to their respective base as follows: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> so the unknown base. . .
At step 438 at least one of the ***ratios*** of ***ratios*** is greater than the upper ***ratio*** cutoff and the unknown base is called as the base complementary to the highest ***ratio*** of ***ratios*** . The code for the base

complementary to the highest ***ratio*** of ***ratios*** would be placed in column 17 and a confidence code of 1 would be placed in column 18 of the. . .

Assume for the purposes of this example that the lower ***ratio*** cutoff is 1.5 and the upper ***ratio*** cutoff is 3. Again, the ***ratios*** of ***ratios*** are as follows: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> As all the ***ratios*** of ***ratios*** are not less than the upper ***ratio*** cutoff, the unknown base is called the base complementary to the highest ***ratio*** of ***ratios*** . The highest ***ratio*** of ***ratios*** is C:C, which has a complementary base G. Thus, the unknown base is called G which is placed in column. . .

The example shows how the unknown base in the sample ***nucleic*** ***acid*** ***sequence*** was correctly called as base G. Although the complementary "mutation" probe associated with the base G (3'-GCCT) did not have the highest fluorescence ***intensity*** , the unknown base was called as base G because the associated "mutation" probe had the highest ***ratio*** increase over the other "mutation" probes.

Fig. . . . level flow of another implementation of the reference method. As in the previous implementation, this implementation also compares the probe ***intensities*** of a reference ***sequence*** to the probe ***intensities*** of a sample ***sequence*** . However, this implementation differs conceptually from the previous implementation in that neighboring probe ***intensities*** are also analyzed, resulting in more accurate base calling.

As a simple example of this implementation of the reference method, suppose a reference ***sequence*** has a ***sequence*** of 5'-AAACCCAATCCACATCA-3' (SEQ ID NO:6) and a sample ***sequence*** has a ***sequence*** of 5'-AAACCCAGTCCACATCA-3' (SEQ ID NO:7), where the mutant base is underlined. Thus, there is a mutation of A to G. Suppose further that the reference and sample ***sequences*** are tiled on chips with the reference ***sequence*** being the chip wild-type. This implementation of the reference method will be described as identifying this mutation base.

For . . . aid the reader in understanding the method. The mutant base position is at position 241 in the reference and sample ***sequences*** , which is shown in bold in the data table.

At step 502 the base ***intensities*** of the reference and sample ***sequences*** are adjusted by subtracting the background or "blank" cell ***intensity*** from each base ***intensity*** . Preferably, if a base ***intensity*** is

then less than or equal to zero, the base ***intensity*** is set equal to a small positive number to prevent division by zero or negative numbers. In the data table, data 502A is the background subtracted base ***intensities*** for the reference ***sequence*** and data 502B is the background subtracted base ***intensities*** for the sample ***sequence*** (also called the "mutant" ***sequence*** in the data table).

At step 504 the base ***intensity*** associated with the reference wild-type is checked to see if it has sufficient ***intensity*** to call the unknown base. In this example, the reference wild-type is base A at position 241. The base ***intensity*** associated with the reference wild-type is identified by a lower case "a" in the left hand column. Thus, the base ***intensities*** in the data table are not identified by their complements and the reference wild-type at the mutation position has an ***intensity*** of 385. The reference wild-type ***intensity*** of 385 is checked by determining if its ***intensity*** is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the ***intensity*** the base ***intensities*** must be over the background ***intensity*** in order to correctly call the unknown base. Thus, the base ***intensity*** associated with the reference wild-type must be greater than the background difference cutoff or the unknown base is not callable.. . .

If the base ***intensity*** associated with the reference wild-type is not greater than the background difference cutoff, the wild-type ***sequence*** would fail to have sufficient ***intensity*** as shown at step 506. Otherwise, at step 508 the wild-type ***sequence*** would pass by having sufficient ***intensity*** .

At step 510 calculations are performed on the background subtracted base ***intensities*** of the reference ***sequence*** in order to "normalize" the ***intensities*** . Each position in the reference ***sequence*** has four background subtracted base ***intensities*** associated with it. The ***ratio*** of the ***intensity*** of each base to the sum of the ***intensities*** of the possible bases (all four) is calculated, resulting in four ***ratios*** , one for each base as shown in the data table. Thus, the following ***ratios*** would be calculated at each position in the reference ***sequence*** : <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> At position 241, A ***ratio*** would be the wild-type ***ratio*** . These ***ratios*** are generally calculated in order to "normalize" the ***intensity*** data as the photon counts may vary widely

from experiment to experiment. Thus, the ***ratios*** provide a way of reconciling the ***intensity*** variations across experiments. Preferably, if the photon counts do not vary widely from experiment to experiment, the probe ***intensities*** do not need to be "normalized."

At step 512 the highest base ***intensity*** associated with the sample ***sequence*** is checked to see if it has sufficient ***intensity*** to call the unknown base. The ***intensity*** is checked by determining if the highest ***intensity*** sample base is greater than the background difference cutoff. If the ***intensity*** is not greater than the background difference cutoff, the sample ***sequence*** fails to have sufficient ***intensity*** as shown at step 514. Otherwise, at step 516 the sample ***sequence*** passes by having sufficient ***intensity*** .

At step 518 calculations are performed on the background subtracted base ***intensities*** of the sample ***sequence*** in order to "normalize" the ***intensities*** . Each position in the sample ***sequence*** has four background subtracted base ***intensities*** associated with it. The ***ratios*** of the ***intensity*** of each base to the sum of the ***intensities*** of the possible bases (all four) are calculated, resulting in four ***ratios*** , one for each base as shown in the data table.

At step 520 if either the reference or sample ***sequences*** failed to have sufficient ***intensity*** , the unknown base is assigned the code N (insufficient ***intensity***) as shown at step 522.

At step 524 the normalized base ***intensities*** of the reference ***sequence*** are subtracted from the normalized base ***intensities*** of the sample ***sequence*** . Thus, at each position the following calculations are performed: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> where the reference and sample ***ratios*** are calculated at steps 510 and 518, respectively. The base differences resulting from these calculations are shown in the data. . .

At . . . a base difference lower than a lower difference cutoff. For example, Fig. 11C shows a graph the normalized sample base ***intensities*** minus the normalized reference base ***intensities*** . Suppose that the upper difference cutoff is 0.15 and the lower difference cutoff is -0.15 as shown by the horizontal. . .

At step 530 the ***ratio*** of the highest background subtracted base ***intensity*** in the sample to the background subtracted reference wild-type base ***intensity*** is calculated. For example, at the mutation position 241 in the

data table, the highest background subtracted base ***intensity*** in the sample is 571 (base G). The background subtracted reference wild-type base ***intensity*** is 385 (base A). The ***ratio*** of 571:385 is calculated and results in 1.48 as shown in the data table.

At step 532 these ***ratios*** are compared to a ***ratio*** at a neighboring position. The ***ratio*** for the n.supth. position is subtracted from the ***ratio*** for the r.supth. position, where r = n + 1. For example, at the mutation position 241 in the data table, the ***ratio*** at position 242 (which equals 1.02) is subtracted from the ***ratio*** at position 241 (which equals 1.48). It has been found that a mutant can be confidently detected by analyzing the difference of these neighboring ***ratios*** .

Fig. . . . 0.28, which is greater than the upper difference cutoff of 0.15. Therefore, the base at position 241 in the sample ***sequence*** is called a base G, which is a mutation from the reference ***sequence*** (A to G).

This second implementation of the reference method is preferable in some instances as it takes into account probe ***intensities*** of neighboring probes. Thus, the first implementation may not have detected the A to G mutation in this example.

The . . . is that the correct base can be called even in the presence of significant levels of cross-hybridization, as long as ***ratios*** of ***intensities*** are fairly consistent from experiment to experiment. In practice, the number of miscalls and ambiguities is significantly reduced, while the. . .

The statistical method is a method of calling bases in a sample ***nucleic*** ***acid*** ***sequence*** . The statistical method utilizes the statistical variation across experiments to call the bases. Therefore, the statistical method is preferable when. . . data from multiple experiments is available and the data is fairly consistent across the experiments. The method compares the probe ***intensities*** of a sample ***sequence*** to statistics of probe ***intensities*** of a reference ***sequence*** in multiple experiments.

For simplicity, the statistical method will be described as being used to identify one unknown base in a sample ***nucleic*** ***acid*** ***sequence*** . In practice, the method is used to identify many or all the bases in a ***nucleic*** ***acid*** ***sequence*** .

The unknown base will be called by comparing the probe ***intensities*** of a sample ***sequence*** to statistics on probe ***intensities*** of a reference ***sequence*** in multiple experiments. Generally, the probe ***intensities*** of the sample ***sequence*** and the reference

sequence experiments are from chips having the same chip wild-type. However, the reference ***sequence*** may or may not be equal to the chip wild-type, as it may have mutations. A base at the same position in the reference and sample ***sequences*** will be associated with up to four mutation probes and a "blank" cell. As before, because each mutation probe is identifiable by the mutation base, the mutation probes' ***intensities*** will be referred to as the "base ***intensities*** " of their respective mutation bases.

As a simple example of the statistical method, suppose a ***gene*** of interest (target) has the ***sequence*** 5'-AAAACTGAAAA-3' (SEQ ID NO:4). Suppose a reference ***sequence*** has the ***sequence*** 5'-AAAACCGAAAA-3' (SEQ ID NO:5), which differs from the target ***sequence*** by the underlined base. Suppose further that a sample ***sequence*** is suspected to have the same ***sequence*** as the target ***sequence*** except for a T base mutation at the underlined base position in 5'-AAAACTGAAAA-3' (SEQ ID NO:4). Suppose that in multiple experiments the reference ***sequence*** is marked and exposed to probes on a chip. Suppose further the sample ***sequence*** is also marked and exposed to probes on a chip. The following are complementary "mutation" probes that could be used for a reference experiment and the sample ***sequence*** : <table> The "mutation" probes shown for the reference ***sequence*** may be from only one experiment, the other experiments may have different "mutation" probes, chip wild-types, tiling methods, and the like. Although each fluorescence ***intensity*** is from a probe, since the probes may be identified by their unique mutation bases, the probe ***intensities*** may be identified by their respective bases as follows: <table> Thus, base A of the reference ***sequence*** will be described as having an ***intensity*** which corresponds to the ***intensity*** of the mutation probe with the mutation base A. The statistical method will now be described as calling the unknown base in the sample ***sequence*** by using this example.

Fig. 12 illustrates the high level flow of the statistical method. At step 602 the four base ***intensities*** associated with the sample ***sequence*** and each of the multiple reference experiments are adjusted by subtracting the background or "blank" cell ***intensity*** from each base ***intensity*** . Preferably, if a base ***intensity*** is then less than or equal to zero, the base ***intensity*** is set equal to a small positive number to prevent division by zero or negative numbers.

At step 604 the ***intensities*** of the reference wild-type bases in the multiple experiments are checked to see if they all

have sufficient ***intensity*** to call the unknown base. The ***intensities*** are checked by determining if the ***intensity*** of the reference wild-type base of an experiment is greater than a predetermined background difference cutoff. The wild-type probe shown earlier for the reference ***sequence*** is 3'-TGGC, and thus the G base ***intensity*** is the wild-type base ***intensity*** . These steps are analogous to steps in the other two methods described herein.

If the ***intensity*** of any one of the reference wild-type bases is not greater than the background difference cutoff, the wild-type experiments fail to have sufficient ***intensity*** as shown at step 606. Otherwise, at step 608 the wild-type experiments pass by having sufficient ***intensity*** .

At step 610 calculations are performed on the background subtracted base ***intensities*** of each of the reference experiments in order to "normalize" the ***intensities*** . Each reference experiment has four background subtracted base ***intensities*** associated with it: one wild-type and three for the other possible bases. In this example, the G base ***intensity*** is the wild-type, the A, C, and T base ***intensities*** being the "other" ***intensities*** . The ***ratios*** of the ***intensity*** of each base to the sum of the ***intensities*** of the possible bases (all four) are calculated, giving one wild-type ***ratio*** and three "other" ***ratios*** . Thus, the following ***ratios*** would be calculated: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> where G ***ratio*** is the wild-type ***ratio*** and A, C, and T ***ratios*** are the "other" ***ratios*** . These four ***ratios*** are calculated for each reference experiment. Thus if the number of reference experiments is n, there would be 4n ***ratios*** calculated. These ***ratios*** are generally calculated in order to "normalize" the ***intensity*** data, as the photon counts may vary widely from experiment to experiment. However, if the probe ***intensities*** do not vary widely from experiment to experiment, the probe ***intensities*** do not need to be "normalized."

At step 612 statistics are prepared for the ***ratios*** calculated for each of the reference experiments. As stated before, each reference experiment will be associated with one wild-type ***ratio*** and three "other" ***ratios*** . The mean and standard deviation are calculated for all the wild-type ***ratios*** . The mean and standard deviation are also calculated for each of the other ***ratios*** , resulting in three other means and standard deviations for each of the bases that is not the wild-type base. Therefore,. . . calculated:

<mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> where the mean and standard deviation of the G ***ratios*** are also known as the wild-type mean and the wild-type standard deviation, respectively. The mean and standard deviation of the. . .

At step 614 the highest base ***intensity*** associated with the sample ***sequence*** is checked to see if it has sufficient ***intensity*** to call the unknown base. The ***intensity*** is checked by determining if the highest ***intensity*** unknown base is greater than the background difference cutoff. If the ***intensity*** is not greater than the background difference cutoff, the sample ***sequence*** fails to have sufficient ***intensity*** as shown at step 616. Otherwise, at step 618 the sample ***sequence*** passes by having sufficient ***intensity*** .

At step 620 calculations are performed on the four background subtracted ***intensities*** of the sample ***sequence*** . The ***ratios*** of the background subtracted ***intensity*** of each base to the sum of the background subtracted ***intensities*** of the possible bases (all four) are calculated, giving four ***ratios*** , one for each base. For consistency, the ***ratio*** associated with the reference wild-type base is called the wild-type ***ratio*** , with there being three "other" ***ratios*** . Thus, the following ***ratios*** are calculated: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> where ***ratio*** G is the wild-type ***ratio*** and ***ratios*** A, C, and T are the "other" ***ratios*** .

Suppose the background subtracted ***intensities*** associated with the sample are as follows:

A -> 310
C -> 50
G -> 26
T -> 100

Then, the corresponding ***ratios*** would be as follows: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> At step 622 if either the reference experiments or the sample ***sequence*** failed to have sufficient ***intensity*** , the unknown base is assigned the code N (insufficient ***intensity***) as shown at step 624.

At step 626 the wild-type and "other" ***ratios*** associated with the sample ***sequence*** are compared to statistical expressions. The statistical expressions include four predetermined standard deviation cutoffs, one associated with each base. Thus,. . . standard deviation cutoffs are set as follows:

<mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> The wild-type base ***ratio*** associated with the sample is compared to a corresponding statistical expression: <mathematical formula> where the WT base std. dev. cutoff is the standard deviation cutoff for the wild-type base. As the wild-type base is G, the above ***comparison*** solves to the following: <mathematical formula> <mathematical formula> <mathematical formula> which is not a true expression (0.05 is not greater. . .

Each of the "other" ***ratios*** associated with the sample is compared to a corresponding statistical expression: <mathematical formula> where the Other base std. dev. cutoff is the standard deviation cutoff for the particular "other" base. Thus, the above ***comparison*** solves to the following three expressions: <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula> <mathematical formula>. . .

At step 628 if only the wild-type ***ratio*** of the sample ***sequence*** was greater than the statistical expression, the unknown base is assigned the code of the reference wild-type base as shown. . .

At step 632 if one or more of the "other" ***ratios*** of the sample ***sequence*** were greater than their respective statistical expressions, the unknown base is assigned an ambiguity code that indicates the unknown base may be any one of the complements of these bases, including the reference wild-type. In this example, the "other" ***ratios*** for A, C, and T were all greater than their corresponding statistical expression. Thus, the unknown base would be called. . .

If none of the ***ratios*** are greater than their respective statistical expressions, the unknown base is assigned the code X (insufficient discrimination) as shown at. . .

The . . . to call the unknown base. The statistical method has also been used to implement confidence estimates and calling of mixed ***sequences*** .

The present invention provides pooling processing which is a method of processing reference and sample ***nucleic*** ***acid*** ***sequences*** together to reduce variations across individual experiments. In the representative embodiment discussed herein, the reference and sample ***nucleic*** ***acid*** ***sequences*** are labeled with different fluorescent markers emitting light at different wavelengths. However, the ***nucleic*** ***acids*** may be labeled with other types of markers including distinguishable radioactive markers.

After the reference and sample ***nucleic*** ***acid***

sequences are labeled with different color fluorescent markers, the labeled reference and sample ***nucleic*** ***acid*** ***sequences*** are then combined and processed together. An apparatus for detecting targets labeled with different markers is provided in U.S. Application. . .

Fig. 13 illustrates the pooling processing of a reference and sample ***nucleic*** ***acid*** ***sequence*** . At step 702 a reference ***nucleic*** ***acid*** ***sequence*** is marked with a fluorescent dye, such as fluorescein. At step 704 a sample ***nucleic*** ***acid*** ***sequence*** is marked with a dye that, upon excitation, emits light of a different wavelength than that of the fluorescent dye of the reference ***sequence*** . For example, the sample ***nucleic*** ***acid*** ***sequence*** may be marked with rhodamine. Alternatively, the sample ***nucleic*** ***acid*** ***sequence*** may be marked by attaching biotin to the sample ***sequence*** which will subsequently bind to streptavidin labeled with phycoerythrin. Of course, either ***sequence*** may be marked with these or other dyes or other kinds of markers (e.g., radioactive) as long as the other ***sequence*** is marked with a marker that is distinguishable. At step 706 the labeled reference ***sequence*** and the labeled sample ***sequence*** are combined. After this step, processing continues in the same manner as for only one labeled ***sequence*** . At step 708 the ***sequences*** are fragmented. The fragmented ***nucleic*** ***acid*** ***sequences*** are then hybridized on a chip containing probes as shown at step 710.

At step 712 a scanner generates image files that indicate the locations where the labeled ***nucleic*** ***acids*** bound to the chip. There is typically some overlap between the two ***signals*** . This is corrected for prior to further analysis, i.e., after correction, the data files correspond to "reference" and "sample." In. . .

The scanner creates an image file for the data associated with each fluorescent marker, indicating the locations where the correspondingly labeled ***nucleic*** ***acid*** bound to the chip. Based upon an analysis of the fluorescence ***intensities*** and locations, it becomes possible to extract information such as the monomer ***sequence*** of ***DNA*** or RNA.

Pooling . . . is common. Although pooling processing has been described as being used to improve the combined processing of reference and sample ***nucleic*** ***acid*** ***sequences*** , the process may also be used for two reference ***sequences*** , two sample ***sequences*** , or multiple ***sequences*** by utilizing multiple distinguishable markers.

Pooling processing may also be utilized with methods of the present invention of identifying mutations in a sample ***nucleic*** ***acid*** ***sequence*** . These methods are highly accurate in identifying single mutations, locating multiple mutations and removing false positives for mutations, where a. . . base calling methods. These methods may be advantageously combined with the base calling methods described herein to efficiently and accurately ***sequence*** a sample ***nucleic*** ***acid*** ***sequence*** .

As discussed earlier in reference to Fig. 8, the fluorescent ***intensities*** of cells near an interrogation position having a mutation are relatively dark which creates "dark regions" around the mutation. These lower fluorescent ***intensities*** result because the cells at interrogation positions near a mutation do not contain probes that are perfectly complementary to the sample ***sequence*** . Thus, the hybridization of these probes with the sample ***sequence*** is lower. The characteristics of these "dark regions" may be utilized to identify mutations and false positives.

For example, a sample ***sequence*** and a reference ***sequence*** were labeled with different fluorescent markers, in this case fluorescein and biotin/phycoerythrin. The sample and reference ***sequences*** are known and the sample ***sequence*** is identical to the reference ***sequence*** except for mutations at certain known positions. The sample and reference ***sequences*** were then processing together using the pooling processing described above and the ***sequences*** were hybridized to a chip including wild-type probes that are perfectly complementary to the reference ***sequence*** . The chip included 20-mer probes with the interrogation position of each probe being at the 12.supth. base position in the. . .

Fig. 14A shows a graph of the scaled fluorescent ***intensities*** (photon counts) of the wild-type probes hybridizing with the sample and reference ***sequences*** . Along the bottom of the graph are numbers which represent wild-type cell positions on the chip. The photon counts of. . .

As shown in Fig. 14A, the scaled photon counts for the wild-type probes hybridizing with the sample and reference ***sequences*** are almost the same except for two "bubbles." A bubble 730 has a top curve defined by the photon counts of the wild-type probes that hybridized with the reference ***sequence*** and a bottom curve defined by the photon counts of the wild-type probes that hybridized with the sample ***sequence*** . Following bubble 730, there is a section 732 where the photon counts for the wild-type probes hybridizing with the sample and reference ***sequences*** are almost the same. After section 732 is another bubble 734 which again has a top curve defined by the

hybridization of the reference ***sequence*** and the bottom curve defined by the hybridization of the sample ***sequence*** . Another partial bubble is shown to the right of bubble 734. Each . . . a dark region surrounding a single mutation. Because the wild-type probes at and surrounding a mutant position in the sample ***sequence*** contain a ***single*** ***base*** ***mismatch*** with the sample ***sequence*** , the hybridization is relatively lower which results in lower photon counts. Much information about the sample ***sequence*** may be acquired by a detailed analysis of these bubble regions. The . . . is believed to be approximately equal to the probe length because each of the probes in this region have a ***single*** ***base*** ***mismatch*** with the sample ***sequence*** .

If . . . example, assume that at wild-type cell number 45 in Fig. 14A, the hybridization of the wild-type probe with the sample ***sequence*** was very low (e.g., around 1000 photon counts). A base calling algorithm that calls the bases according to the ***intensities*** among the cells at that position may indicate that there is a mutation at this position. However, the low photon. . .

Additionally, . . . eliminated because they are incompatible with the bubble size (which indicates single mutation, for example). Also, by identifying clearly a " ***mismatch*** zone," we can now apply algorithms that factor in the effect of a ***mismatch*** or multiple mismatches.

Additionally, the shape of the bubble may indicate what mutation has occurred. Fig. 14B shows a hypothetical graph of the fluorescent ***intensities*** vs. cell locations for wild-type probes hybridizing with two sample ***sequences*** and one reference ***sequence*** . A C-A ***mismatch*** will be more destabilizing to probe hybridization than a U-G ***mismatch*** . As shown, the more destabilizing C-A ***mismatch*** results in a larger volume bubble. The shape of the bubble may be utilized to identify the particular mutation by. . .

Fig. 14C shows a graph of the fluorescent ***intensities*** (photon counts) of the wild-type probes hybridizing with the sample and reference ***sequences*** . A single bubble 750 is flanked on either side by regions 752 and 754 which do not contain a mutation.. . .

Fig. . . . the present invention that uses the hybridization data from more than one base position to identify mutations in a sample ***nucleic*** ***acid*** ***sequence*** . After probe ***intensities*** from the hybridization of wild-type probes with a sample and reference ***sequence*** are

measured, the system identifies a bubble region at step 780. Bubble regions are identified as regions where the hybridization of the wild-type probes to the sample and reference ***sequence*** differ significantly. Additionally, the reference ***sequence*** should hybridize more strongly with the wild-type probes since the wild-type probes will be perfectly complementary to the reference ***sequence*** .

At . . . if they are approximately equal. If the bubble width is approximately equal to the probe length, the bubble represents a ***single*** ***base*** mutation at step 788. Again, the determination of how close the bubble width should be to the probe length may. . .

If . . . their positions by analyzing the pattern of the bubble. The base calling method with the present invention may be the ***intensity*** ***ratio*** method, reference method, statistical method, or any other method.

Although in a preferred embodiment, this method of identifying mutations in a sample ***nucleic*** ***acid*** ***sequence*** is utilized in conjunction with pooling processing in order to reduce variations, the method may be utilized without pooling processing.. . .

The present invention provides a method of comparative analysis and visualization of multiple experiments. The method allows the ***intensity*** ***ratio*** , reference, and statistical methods to be run on multiple datafiles simultaneously. This permits different experimental conditions, sample preparations, and analysis parameters to be compared in terms of their effects on ***sequence*** calling. The method also provides verification and editing functions, which are essential to reading ***sequences*** , as well as navigation and analysis tools.

Fig. . . . ID NO:8 and SEQ ID NO:9). The windows shown are from an appropriately programmed Sun Workstation. However, the comparative analysis ***software*** may also be implemented on or ported to a personal ***computer*** , including IBM PCs and compatibles, or other workstation environments. A window 802 is shown having pull down menus for the. . .

The main section of the window is divided into a reference ***sequence*** area 814 and a sample ***sequence*** area 816. The reference ***sequence*** area is where known ***sequences*** are displayed and is divided into a reference name subarea 818 and reference base subarea 820. The reference name subarea is shown with the filenames that contain the reference ***sequences*** . The chip wild-type is identified by the filename with the extension ".wt#" where the # indicates a unit on the chip. The reference base subarea contains the bases of the reference ***sequences*** . A capital C 822 is displayed to the right of the reference ***sequence*** that is the chip

wild-type for the current analysis. Although the chip wild-type ***sequence*** has associated fluorescence ***intensities*** , the other reference ***sequences*** shown below the chip wild-type may be known ***sequences*** that have not been tiled on the chip. These may or may not have associated fluorescence ***intensities*** . The reference ***sequences*** other than the chip wild-type are used for ***sequence*** ***comparisons*** and may be in the form of simple ASCII text files.

Sample ***sequence*** area 816 is where sample or unknown experimental ***sequences*** are displayed for ***comparison*** with the reference ***sequences*** . The sample ***sequence*** area is divided into a sample name subarea 824 and sample base subarea 826. The sample name subarea is shown with filenames that contain the sample ***sequences*** . The filename extensions indicate the method used to call the sample ***sequence*** where ".cq#" denotes the ***intensity*** ***ratio*** method, ".rq#" denotes the reference method, and ".sq#" denotes the statistical method (# indicates the unit on the chip). The sample base subarea contains the bases of the sample ***sequences*** . The bases of the sample ***sequences*** are identified by the codes previously set forth which, for the most part, conform to the IUPAC standard. Window . . . and the pathname of the file containing the base is displayed in the message panel. The base's position in the ***nucleic*** ***acid*** ***sequence*** is also displayed in the message panel.

In pull down menu File 804, the user is able to load files of experimental ***sequences*** that have been tiled and scanned on a chip. There is a chip wild-type associated with each experimental ***sequence*** . The chip wild-type associated with the first experimental ***sequence*** loaded is read and shown as the chip wild-type in reference ***sequence*** area 814. The user is also able to load files of known ***nucleic*** ***acid*** ***sequences*** as reference ***sequences*** for ***comparison*** purposes. As before, these known reference ***sequences*** may or may not have associated probe ***intensity*** data. Additionally, in this menu the user is able to save ***sequences*** that are selected on the screen into a project file that can be loaded in at a later time. The project file also contains any linkage of the ***sequences*** , where ***sequences*** are linked for ***comparison*** purposes. ***Sequences*** to be saved, both reference and sample, are chosen by selecting the ***sequence*** filename with an input device in the reference or sample name subareas.

In pull down menu Edit 806, the user is able to link together ***sequences*** in the reference and sample ***sequence***

areas. After the user has selected one reference and one or more sample ***sequences*** , the sample ***sequences*** can be linked to the reference ***sequence*** by selecting an entry in the pull down menu. Once the ***sequences*** are linked, a link number 830 is displayed next to each of ***sequences*** of related interest. Each group of linked ***sequences*** is associated with a unique link number, so the user can easily identify which ***sequences*** are linked together. Linking ***sequences*** permits the user to more easily compare the linked ***sequences*** . The user is also able to remove and display links from this menu.

In pull down menu View 808, the user is able to display ***intensity*** graphs for selected bases. Once a base is selected in the reference or sample base subareas, the user may request an ***intensity*** graph showing the hybridized probe ***intensities*** of the selected base and a delineated neighborhood of bases near the selected base. ***Intensity*** graphs may be displayed for one or multiple selected bases. The user is also able to prepare comment files and. . .

Fig. 17 illustrates an ***intensity*** graph window for a selected base at position 120 (SEQ ID NO:30 and SEQ ID NO:31). The filename containing the ***sequence*** data is displayed at 904. The graph shows the ***intensities*** for each of the hybridized probes associated with a base. Each grouping of four vertical bars on the graph, which are labeled as "a", "c", "g", and "t" on line 906, shows the background subtracted ***intensities*** of probes having the indicated substitution base. In one embodiment, the called bases are shown in red. The wild-type base. . . called base is M which means the base is either A or C (amino). The user is able to use ***intensity*** graphs to visually compare the ***intensities*** of each of the possible calls.

Fig. 18 illustrates multiple ***intensity*** graph windows for selected bases (SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35). There are three ***intensity*** graph windows 1002, 1004, and 1006 as shown. Each window may be associated with a different experiment, where the ***sequence*** analyzed in the experiment may be either a reference (if it has associated probe ***intensity*** data as in the chip wild-type) or a sample ***sequence*** . The windows are aligned and a rectangular box 1008 shows the selected bases' position in each of the ***sequences*** (position 162 in Fig. 18). The rectangular box aids the user in identifying the selected bases.

Referring again to Fig. 16, in pull down menu Highlight 810, the user is able to compare the ***sequences*** of references and samples. At least four ***comparisons*** are available to the user, including the following: sample ***sequences*** to the

chip wild-type ***sequence*** , sample ***sequences*** to any reference ***sequences*** , sample ***sequences*** to any linked reference ***sequences*** , and reference ***sequences*** to the chip wild-type ***sequence*** . For example, after the user has linked a reference and sample ***sequence*** , the user can compare the bases in the linked ***sequences*** . Bases in the sample ***sequence*** that are different from the reference ***sequence*** will then be indicated on the display device to the user (e.g., base is shown in a different color). In another example, the user is able to perform a ***comparison*** that will help identify sample ***sequences*** . After a sample is linked to multiple reference ***sequences*** , each base in the sample ***sequence*** that does not match the wild-type ***sequence*** is checked to see if it matches one of the linked reference ***sequences*** . The bases that match a linked reference ***sequence*** will then be indicated on the display device to the user. The user may then more easily identify the sample ***sequence*** as being one of the reference ***sequences*** .

In pull down menu Help 812, the user is able to get information and instructions regarding the comparative analysis ***program*** , the calling methods, and the IUPAC definitions used in the ***program*** .

Fig. 19 illustrates the ***intensity*** ***ratio*** method correctly calling a mutation in solutions with varying concentrations (SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ. . . SEQ ID NO:17, and SEQ ID NO:18). A window 1102 is shown with a chip wild-type 1104 and a mutant ***sequence*** 1106. The mutant ***sequence*** differs from the chip wild-type at the position indicated by the rectangular box 1108. The chip wild-type and mutant ***sequences*** are a region of HIV Pol ***Gene*** spanning mutations occurring in AZT drug therapy. There are seven sample ***sequences*** that are called using the ***intensity*** ***ratio*** method. The sample ***sequences*** are actually solutions of different proportions of the chip wild-type ***sequence*** and the mutant ***sequence*** . Thus, there are sample solutions 1110, 1112, 1114, 1116, 1118, 1120, and 1122. The solutions are 15-mer tilings across the chip wild-type with increased percentages of the mutant ***sequence*** from 0 to 100% by weight. The following shows the proportions of the sample solutions: <table> For example, sample solution 1114 contains 75% chip wild-type ***sequence*** and 25% mutant ***sequence*** .

Now referring to the bases called in rectangular box 1108 for the sample solutions, the ***intensity*** ***ratio*** method correctly calls sample solution 1110 as having a base A as in the chip-wild type ***sequence*** . This is correct because sample

solution 1110 is 100% chip wild-type ***sequence*** . The ***intensity*** ***ratio*** method also calls sample solution 1112 as having a base A because the sample solution is 90% chip wild-type ***sequence*** .

The ***intensity*** ***ratio*** method calls the identified base in sample solutions 1114 and 1116 as being an R, which is an ambiguity IUPAC. . . or G (purine). This also a correct base call because the sample solutions have from 75% to 50% chip-wild type ***sequence*** and from 25% to 50% mutation ***sequence*** . Thus, the ***intensity*** ***ratio*** method correctly calls the base in this transition state.

Sample solutions 1118, 1120, and 1122 are called by the ***intensity*** ***ratio*** method as having a mutation base G at the specified location. This is a correct base call because the sample solutions primarily consist of the mutation ***sequence*** (75%, 90%, and 100% respectively). Again, the ***intensity*** ***ratio*** method correctly called the bases.

These . . . show that the base calling methods of the present invention may also be used for solutions of more than one ***nucleic*** ***acid*** ***sequence*** .

Fig. 20 illustrates the reference method correctly calling a mutant base where the ***intensity*** ***ratio*** method incorrectly called the mutant base (SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, and SEQ ID NO:39). There are three ***intensity*** graph windows 1202, 1204, and 1206 as shown. The windows are aligned and a rectangular box 1208 outlines the bases of interest.

Window 1202 shows a sample ***sequence*** called using the ***intensity*** ***ratio*** method. However, the base in the rectangular box 1208 was incorrectly called base C, as there is actually a base A at that position. The ***intensity*** ***ratio*** method incorrectly called the base as C because the probe ***intensity*** associated with base C is much higher than the other probe ***intensities*** .

Window 1204 shows a reference ***sequence*** called using the ***intensity*** ***ratio*** method. As the reference ***sequence*** is known, it is not necessary to know the method used to call the reference ***sequence*** . However, it is important to have probe ***intensities*** for a reference ***sequence*** to use the reference method. The reference ***sequence*** is called a base C at the position indicated by the rectangular box.

Window 1206 shows the sample ***sequence*** called using the reference method. The reference method correctly calls the specified base as being base A. Thus, for some cases the reference method is preferable to the ***intensity*** ***ratio*** method because it compares probe ***intensities*** of a sample

sequence to probe ***intensities*** of a reference ***sequence*** .

The ***intensity*** ***ratio*** method was used in ***sequence*** analysis of various polymorphic HIV-1 clones using a protease chip. Single stranded ***DNA*** of a 382 nt region was used with 4 different clones (HXB2, SF2, NY5, pPol4mut18). Results were compared to results from an ABI ***sequencer*** . The results are illustrated below: <table>

HIV protease genotyping was performed using the described chips and CallSeq.trade. ***intensity*** ***ratio*** calculations. Samples were evaluated from AIDS patients before and after ddI treatment. Results were confirmed with ABI ***sequencing*** .

Fig. 21 illustrates the output of the ViewSeq.trade. ***program*** with four pretreatment samples and four posttreatment samples (SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25,. . .

The . . . of this disclosure. Merely by way of example, while the invention is illustrated with particular reference to the evaluation of ***DNA*** (natural or unnatural), the methods can be used in the analysis from chips with other materials synthesized thereon, such as. . .

CLMEN 1. In a ***computer*** system, a method of identifying an unknown base in a sample ***nucleic*** ***acid*** sequence, said method comprising the steps of:

inputting a plurality of probe ***intensities*** , each of said probe ***intensities*** being associated with a ***nucleic*** ***acid*** probe;

said ***computer*** system comparing said plurality of probe ***intensities*** wherein each of said plurality of probe ***intensities*** is substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with at least one ***nucleic*** ***acid*** sequence, said at least one ***nucleic*** ***acid*** sequence including said sample sequence; and

calling said unknown base according to results of said comparing step.

2. In a ***computer*** system, a method of identifying an unknown base in a sample ***nucleic*** ***acid*** sequence, said method comprising the steps of:

inputting a plurality of probe ***intensities*** , each of said probe ***intensities*** being associated with a ***nucleic*** ***acid*** probe;

said ***computer*** system comparing said plurality of probe ***intensities*** wherein each of said plurality of probe ***intensities*** is substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with

said sample sequence; and

calling said unknown base according to results of said comparing step.

3. The method of claim 2, wherein said comparing step includes the step of said ***computer*** system calculating a ***ratio*** of a higher probe ***intensity*** to a lower probe ***intensity*** .

4. . . . said calling step includes the step of calling said unknown base according to said probe associated with said higher probe ***intensity*** if said ***ratio*** is greater than a predetermined ***ratio*** value.

5. The method of claim 4, wherein said predetermined ***ratio*** value is approximately 1.2.

6. In a ***computer*** system, a method of identifying an unknown base in a sample ***nucleic*** ***acid*** sequence, said method comprising the steps of:

inputting a first set of probe ***intensities*** , each of said probe ***intensities*** in said first set being associated with a ***nucleic*** ***acid*** probe and substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with a reference ***nucleic*** ***acid*** sequence;

inputting a second set of probe ***intensities*** , each of said probe ***intensities*** in said second set being associated with a ***nucleic*** ***acid*** probe and substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with said sample sequence;

said ***computer*** system comparing at least one of said probe ***intensities*** in said first set and at least one of said probe ***intensities*** in said second set; and

calling said unknown base according to results of said comparing step.

7. The method of claim 6, wherein said comparing step includes the steps of:

calculating first ***ratios*** of a wild-type probe ***intensity*** to each probe ***intensity*** of a probe hybridizing with said reference sequence, wherein said wild-type probe ***intensity*** is associated with a wild-type probe; and

calculating second ***ratios*** of the highest probe ***intensity*** of a probe hybridizing with said sample sequence to each probe ***intensity*** of a probe hybridizing with said sample sequence.

8. The method of claim 7, wherein said comparing step further includes the step of calculating third ***ratios*** of said first ***ratios*** to said second ***ratios*** .

9. . . . said calling step includes the step of calling said

unknown base according to said probe associated with a highest third ***ratio*** .

10. The method of claim 6, wherein said comparing step includes the step of calculating a ***ratio*** of a highest probe ***intensity*** in said first set to a highest ***intensity*** in said second set.

11. The method of claim 10, wherein said comparing step further includes the step of comparing said ***ratio*** of neighboring ***nucleic*** ***acid*** probes.

12. In a ***computer*** system, a method of identifying an unknown base in a sample ***nucleic*** ***acid*** sequence, said method comprising the steps of:

inputting statistics about a plurality of experiments, each of said experiments producing probe ***intensities*** each being associated with a ***nucleic*** ***acid*** probe and substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with a reference ***nucleic*** ***acid*** sequence;

inputting a plurality of probe ***intensities*** , each of said plurality of probe ***intensities*** being associated with a ***nucleic*** ***acid*** probe and substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with said sample sequence;

said ***computer*** system comparing at least one of said plurality of probe ***intensities*** with said statistics; and

calling said unknown base according to results of said comparing step.

15. A method of processing first and second ***nucleic*** ***acid*** sequences, comprising the steps of:

providing a plurality of ***nucleic*** ***acid*** probes;

labeling said first ***nucleic*** ***acid*** sequence with a first marker;

labeling said second ***nucleic*** ***acid*** sequence with a second marker; and

hybridizing said first and second labeled ***nucleic*** ***acid*** sequences at the same time.

16. The method of claim 15, wherein said plurality of ***nucleic*** ***acid*** probes are on a chip.

17. The method of claim 15, further comprising the step of fragmenting said first and second ***nucleic*** ***acid*** sequences at the same time.

18. . . . further comprising the step of scanning for said first and second markers on said chip, said first and second labeled ***nucleic*** ***acid*** sequences being on said chip.

20. In a ***computer*** system, a method of identifying mutations in a sample ***nucleic*** ***acid*** sequence, said method comprising the steps of:

inputting a first set of probe ***intensities*** , each of said probe ***intensities*** in said first set being associated with a ***nucleic*** ***acid*** probe and substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with a reference ***nucleic*** ***acid*** sequence;

inputting a second set of probe ***intensities*** , each of said probe ***intensities*** in said second set being associated with a ***nucleic*** ***acid*** probe and substantially proportional to said associated ***nucleic*** ***acid*** probe hybridizing with said sample sequence;

said ***computer*** system comparing probe ***intensities*** in said first set and probe ***intensities*** in said second set to select hybridization regions where said probe ***intensities*** in said first set and said probe ***intensities*** in said second set differ; and

identifying mutations according to characteristics of said selected regions.

21. The method of claim 20, wherein said selected regions are determined by comparing probe ***intensities*** of wild-type probes.

23. . . . selected region;

identifying a likely position of a mutation in said selected region according to an interrogation position of said ***nucleic*** ***acid*** probes; and

performing base calling at said likely position.

24. In a ***computer*** system, a method of analyzing a plurality of sequences of bases, said plurality of sequences including at least one reference. . .

25. The method of claim 24, wherein said plurality of sequences are monomer strands of ***DNA*** or RNA.

28. The method of claim 26, further comprising the step of displaying a ***label*** in said first area to identify said chip wild-type sequence.

L332 ANSWE

TIEN ***Computer*** -aided visualization and analysis system for ***nucleic*** ***acid*** sequence evaluation.

TIDE Rechnergestuetztes Anzeigesystem und Analysesystem fuer Nukleinsaeure-Sequenzauswertung.

TIFR Systeme de visualisation et d'analyse assiste par ordinateur pour l'evaluation de sequences d'acides nucleiques.

IN Chee, Mark S., 3199 Waverly Street, Palo Alto, California 94306, US;
Wang, Chunwei, 20350 Stevens Creek Boulevard no. 307, Cupertino, California 95014, US;
Jevons, Luis C., 701 Ramona Avenue, Sunnyvale, California 94087, US;
Bernhart, Derek H., 111 Seale Avenue, Palo Alto, California 94301, US;
Lipshutz, Robert J., 970 Palo Alto Avenue, Palo Alto, California 94301, US

PA AFFYMAX TECHNOLOGIES N.V., De Ruyterkade 62, Willemstad, Curacao, AN

PAN 1316841

AG Nash, David Allan, Haseltine Lake & Co. Hazlitt House 28 Southampton Buildings Chancery Lane, London WC2A 1AT, GB

AGN 59252

OS ESP1996032 EP 0717113 A2 960619

SO Wila-EPZ-1996-H25-T1a

DT Patent

LA Anmeldung in Englisch; Veroeffentlichung in Englisch

DS R DE; R FR; R GB; R IT; R NL

PIT EPA2 EUROPAEISCHE PATENTANMELDUNG

PI EP 717113 A2 960619

OD 960619

AI EP 95-307476 951020

PRAI US 94-327525 941021

ABEN A ***computer*** system (1) for analyzing ***nucleic*** ***acid*** sequences is provided. The ***computer*** system is used to perform multiple methods for determining unknown bases by analyzing the fluorescence ***intensities*** of hybridized ***nucleic*** ***acid*** probes. The results of individual experiments may be improved by processing ***nucleic*** ***acid*** sequences together. Comparative analysis of multiple experiments is also provided by displaying reference sequences in one area (814) and sample sequences in another area (816) on a display device (3).

TIEN ***Computer*** -aided visualization and analysis system for ***nucleic*** ***acid*** sequence evaluation.

A ***computer*** system (1) for analyzing ***nucleic***

acid sequences is provided. The ***computer*** system is used to perform multiple methods for determining unknown bases by analyzing the fluorescence ***intensities*** of hybridized ***nucleic*** ***acid*** probes. The results of individual experiments may be improved by processing ***nucleic*** ***acid*** sequences together. Comparative analysis of multiple experiments is also provided by displaying reference sequences in one area (814) and sample. . .

DETDEN The present invention relates to the field of ***computer*** systems. More specifically, the present invention relates to ***computer*** systems for visualizing biological ***sequences*** , as well as for evaluating and comparing biological ***sequences*** .

Devices and ***computer*** systems for forming and using ***arrays*** of materials on a substrate are known. For example, PCT applications WO92/10588 and 95/11995, incorporated herein by reference for all purposes, describe techniques for ***sequencing*** or ***sequence*** checking ***nucleic*** ***acids*** and other materials. ***Arrays*** for performing these operations may be formed in ***arrays*** according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent Nos. 5,445,934 and 5384,261, and U.S.. . .

According to one aspect of the techniques described therein, an ***array*** of ***nucleic*** ***acid*** probes is fabricated at known locations on a chip or substrate. A labeled ***nucleic*** ***acid*** is then brought into contact with the chip and a scanner generates an image file (also called a cell file) indicating the locations where the labeled ***nucleic*** ***acids*** bound to the chip. Based upon the image file and identities of the probes at specific locations, it becomes possible to extract information such as the monomer ***sequence*** of ***DNA*** or RNA. Such systems have been used to form, for example, ***arrays*** of ***DNA*** that may be used to study and detect mutations relevant to cystic fibrosis, the P53 ***gene*** (relevant to certain cancers), HIV, and other genetic characteristics.

Improved ***computer*** systems and methods are needed to evaluate, analyze, and process the vast amount of information now used and made available. . .

An improved ***computer*** -aided system for visualizing and determining the ***sequence*** of ***nucleic*** ***acids*** is disclosed. The ***computer*** system provides, among other things, improved methods of analyzing fluorescent image files of a chip containing hybridized ***nucleic*** ***acid*** probes in order to call bases in sample ***nucleic*** ***acid*** ***sequences*** .

According to one aspect of the invention, a ***computer***

system is used to identify an unknown base in a sample ***nucleic*** ***acid*** ***sequence*** by the steps of:

inputting multiple probe ***intensities*** , each of the probe ***intensities*** being associated with a ***nucleic*** ***acid*** probe;

the ***computer*** system comparing the multiple probe ***intensities*** where each of the probe ***intensities*** is substantially proportional to a ***nucleic*** ***acid*** probe hybridizing with at least one ***nucleic*** ***acid*** ***sequence*** ; and

calling the unknown base according to the results of the ***comparison*** of the multiple probe ***intensities*** .

According to one specific aspect of the invention, a higher probe ***intensity*** is compared to a lower probe ***intensity*** to call the unknown base. According to another specific aspect of the invention, probe ***intensities*** of a sample ***sequence*** are compared to probe ***intensities*** of a reference ***sequence*** . According to yet another specific aspect of the invention, probe ***intensities*** of a sample ***sequence*** are compared to statistics about probe ***intensities*** of a reference ***sequence*** from multiple experiments.

According to another aspect of the invention, a method is disclosed of processing reference and sample ***nucleic*** ***acid*** ***sequences*** to reduce the variations between the experiments by the steps of:

providing a plurality of ***nucleic*** ***acid*** probes;

labeling the reference ***nucleic*** ***acid*** ***sequence*** with a first marker;

labeling the sample ***nucleic*** ***acid*** ***sequence*** with a second marker; and

hybridizing the labeled reference and sample ***nucleic*** ***acid*** ***sequences*** at the same time.

According to another aspect of the invention, a ***computer*** system is used to identify mutations in a sample ***nucleic*** ***acid*** ***sequence*** by the steps of:

inputting a first set of probe ***intensities*** , each of the probe ***intensities*** in said fisrt set being associated with a ***nucleic*** ***acid*** probe and substantially proportional to the associated ***nucleic*** ***acid*** probe hybridizing with a reference ***nucleic*** ***acid*** ***sequence*** ;

inputting a second set of probe ***intensities*** , each of the probe ***intensities*** in said fisrt set being associated with a ***nucleic*** ***acid*** probe and substantially proportional to the associated ***nucleic*** ***acid***

probe hybridizing with said sample ***sequence*** ;

the ***computer*** system comparing probe ***intensities*** in the first set to probe ***intensities*** in the second set to select hybridization regions where the probe ***intensities*** in the first and second sets differ; and

identifying mutations according to characteristics of the selected regions.

According to yet another aspect of the invention, a ***computer*** system is used for comparative analysis and visualization of multiple ***sequences*** by the steps of:

displaying at least one reference ***sequence*** in a first area on a display device; and

displaying at least one sample ***sequence*** in a second area on said display device;

whereby a user is capable of visually comparing the multiple ***sequences*** .

Fig. 1 illustrates an example of a ***computer*** system used to execute the ***software*** of the present invention;

Fig. 2 shows a system block diagram of a typical ***computer*** system used to execute the ***software*** of the present invention;

Fig. 3 illustrates an overall system for forming and analyzing ***arrays*** of biological materials such as ***DNA*** or RNA;

Fig. 4 is an illustration of the ***software*** for the overall system;

Fig. 5 illustrates the global layout of a chip formed in the overall system;

. . . lanes on a chip;

Fig. 8 illustrates a hybridization pattern of a target on a chip with a reference ***sequence*** as in Fig. 7;

Fig. 9 illustrates the high level flow of the ***intensity*** ***ratio*** method;

Fig. 10A illustrates the high level flow of one implementation of the reference method and Fig. 10B shows. . . shows a data table for use with the reference method; Fig. 11C shows a graph of the normalized sample base ***intensities*** minus the normalized reference base ***intensities*** ; and Fig. 11D shows other graphs of data in the data table;

Fig. 12 illustrates the high level flow of the statistical method;

Fig. 13 illustrates the pooling processing of a reference and sample ***nucleic*** ***acid*** ***sequence*** ;

Figs. 14A and 14C show graphs of scaled fluorescent ***intensities*** of wild-type probes hybridizing with sample and reference ***sequences*** and 14B shows a hypothetical graph of fluorescent ***intensities*** of wild-type probes

hybridizing with two sample ***sequences*** and a reference ***sequence*** ;

Fig. 15 illustrates the high level flow of an embodiment that uses the hybridization data from than one base position to identify mutations in a sample ***sequence*** ;

Fig. 16 illustrates the main screen and the associated pull down menus for comparative analysis and visualization of multiple experiments;

Fig. 17 illustrates an ***intensity*** graph window for a selected base;

Fig. 18 illustrates multiple ***intensity*** graph windows for selected bases;

Fig. 19 illustrates the ***intensity*** ***ratio*** method correctly calling a mutation in solutions with varying concentrations;

Fig. 20 illustrates the reference method correctly calling a mutant base where the ***intensity*** ***ratio*** method incorrectly called the mutant base; and

Fig. 21 illustrates the output of the ViewSeq.trade. ***program*** with four pretreatment samples and four posttreatment samples.

I. General

II. ***Intensity*** ***Ratio*** Method

III. Reference Method

IV. Statistical Method

V. Pooling Processing

VI. Comparative Analysis

VII. Examples. . .

In . . . the systems and methods of the present invention may be advantageously applied to a variety of systems, including IBM personal ***computers*** running MS-DOS or Microsoft Windows. Therefore, the following description of specific systems are for purposes of illustration and not limitation.. . .

Fig. 1 illustrates an example of a ***computer*** system used to execute the ***software*** of the present invention. Fig. 1 shows a ***computer*** system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have. . . houses a floppy disk drive 14 and a hard drive (not shown) that may be utilized to store and retrieve ***software*** ***programs*** incorporating the present invention. Although a floppy disk 15 is shown as the removable media, other removable tangible media including CD-ROM, flash memory and tape may be utilized. Cabinet 7 also houses familiar ***computer*** components (not shown) such as a processor, memory, and the like. Fig. 2 shows a system block diagram of ***computer*** system 1

used to execute the ***software*** of the present invention. As in Fig. 1, ***computer*** system 1 includes monitor 3 and keyboard 9. ***Computer*** system 1 further includes subsystems such as a central processor 52, system memory 54, I/O controller 56, display adapter 58,. . . 64 is representative of an internal hard drive, floppy drive, CD-ROM, flash memory, tape, or any other storage medium. Other ***computer*** systems suitable for use with the present invention may include additional or fewer subsystems. For example, another ***computer*** system could include more than one processor 52 (i.e., a multi-processor system) or memory cache.

Arrows such as 70 represent the system bus architecture of ***computer*** system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, speaker 68. . . could be connected to the other subsystems through a port or have an internal direct connection to central processor 52. ***Computer*** system 1 shown in Fig. 2 is but an example of a ***computer*** system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will. . .

The VLSIPS.trade. technology provides methods of making very large ***arrays*** of ***oligonucleotide*** probes on very small chips. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated by reference for all purposes. The ***oligonucleotide*** probes on the ***DNA*** probe ***array*** are used to detect complementary ***nucleic*** ***acid*** ***sequences*** in a sample ***nucleic*** ***acid*** of interest (the "target" ***nucleic*** ***acid***).

The present invention provides methods of analyzing hybridization ***intensity*** files for a chip containing hybridized ***nucleic*** ***acid*** probes. In a representative embodiment, the files represent fluorescence data from a biological ***array*** , but the files may also represent other data such as radioactive ***intensity*** data or large molecule detection data. Therefore, the present invention is not limited to analyzing fluorescent measurements of hybridizations but. . .

For purposes of illustration, the present invention is described as being part of a ***computer*** system that designs a chip mask, synthesizes the probes on the chip, ***labels*** the ***nucleic*** ***acids*** , and scans the hybridized ***nucleic*** ***acid*** probes. Such a system is fully described in U.S. Patent Application No. 08/249,188 which has been incorporated by reference for. . .

Fig. 3 illustrates a computerized system for forming and analyzing ***arrays*** of biological materials such as RNA or ***DNA***

```
=> s (SBH or sequenc###(p)array#)

        218    FILE AGRICOLA
         40    FILE AIDSLINE
         31    FILE ANABSTR
        113    FILE AQUASCI
         45    FILE BIOBUSINESS
       2048*   FILE BIOSIS
        244*   FILE BIOTECHABS
        244*   FILE BIOTECHDS
        456    FILE CABA
        399    FILE CANCERLIT
       2696    FILE CAPLUS
         37*   FILE CEABA
         17    FILE CEN
         19*   FILE CIN
        915    FILE CJACS
         30    FILE CJELSEVIER
         14    FILE CONFSCI
  18 FILES SEARCHED...
          6    FILE CROPU
          1    FILE DDFB
         15    FILE DDFU
        481    FILE DGENE
        746    FILE DISSABS
          1    FILE DRUGB
         28    FILE DRUGU
         40    FILE EMBAL
       1575    FILE EMBASE
         22*   FILE FSTA
         80    FILE GENBANK
  31 FILES SEARCHED...
          3    FILE HEALSAFE
       4819    FILE IFIPAT
        123    FILE JICST-EPLUS
          4    FILE JPNEWS
        981    FILE LIFESCI
       1884    FILE MEDLINE
        728*   FILE NTIS
         41    FILE OCEAN
         12    FILE PHIN
        511    FILE PROMT
       1913    FILE SCISEARCH
        177    FILE TOXLINE
       1154    FILE TOXLIT
```

```
  45 FILES SEARCHED...
     14206   FILE USPATFULL
       200   FILE NLDB
         8   FILE PNI
         3*  FILE APIPAT
       172   FILE DPCI
      2571   FILE EUROPATFULL
        45*  FILE INPADOC
       375   FILE JAPIO
         6   FILE PAPERCHEM2
         0*  FILE PATDPA
       400   FILE PATOSEP
       151   FILE PATOSWO
        30   FILE PIRA
  62 FILES SEARCHED...
        17*  FILE RAPRA
       118   FILE TULSA
         1   FILE TULSA2
      2439   FILE WPIDS
      2439   FILE WPINDEX

  58 FILES HAVE ONE OR MORE ANSWERS,   67 FILES SEARCHED IN STNINDEX

L1   QUE (SBH OR SEQUENC###(P) ARRAY#)

=> d rank

F1       14206   USPATFULL
F2        4819   IFIPAT
F3        2696   CAPLUS
F4        2571   EUROPATFULL
F5        2439   WPIDS
F6        2439   WPINDEX
F7        2048*  BIOSIS
F8        1913   SCISEARCH
F9        1884   MEDLINE
F10       1575   EMBASE
F11       1154   TOXLIT
F12        981   LIFESCI
F13        915   CJACS
F14        746   DISSABS
F15        728*  NTIS
F16        511   PROMT
F17        481   DGENE
F18        456   CABA
F19        400   PATOSEP
F20        399   CANCERLIT
```

4 ANSWER 16 OF 61 COPYRIGHT 1996 INFO. ACCESS CO.

AN 94:340086 NLDB
TI Chip Deciphers Human Genome
SO Applied Genetics News, (Oct 1994) Vol. 15, No. 3.
ISSN: 0271-7107.
PB Business Communications Company, Inc
DT Newsletter
LA English
WC 333
TX The "super chip" technology is based on ***sequencing*** -by- ***hybridization*** (SBH). The first step is to break down DNA into many pieces, after which the pieces are separated and then. . . unknown DNA in complementary fashion, essentially matching up with defined sequences. By detecting the tags and analyzing the results with ***computer*** software, the base sequence of the unknown DNA can be ascertained (see illustration).

L64 ANSWER 17 OF 61 COPYRIGHT 1996 INFO. ACCESS CO.

AN 94:343793 NLDB
TI HYSEQ GAINS RIGHTS TO "SUPER CHIP" FOR DECODING HUMAN GENOME
SO Biotech Business, (Nov 1994) Vol. 7, No. 11.
ISSN: 0899-5702.
PB Worldwide Videotex
DT Newsletter
LA English
WC 912
TX The . . . Chicago; the university operates Argonne for DOE. Under the agreement, Hyseq was granted exclusive patent rights to a variation of ***Sequencing*** -By- ***Hybridization*** (SBH) technology known as Format 3 that allows large-scale gene sequencing to be done on a "super chip" a 1-inch-square. . .

Sequencing -By- ***Hybridization*** Technology

Sequencing -By- ***Hybridization*** is a revolutionary technology designed to decipher the human genome, all of the genes that comprise the human genetic code.

Using . . . Researchers are able to tell which probes have attached by detecting the tags, and the results are analyzed using special ***computer*** software.

L64 ANSWER 18 OF 61 COPYRIGHT 1996 INFO. ACCESS CO.

AN 94:106336 NLDB
TI ***Computer*** Speeds Genome Mapping
SO Applied Genetics News, (Mar 1994) Vol. 14, No. 8.
ISSN: 0271-7107.
PB Business Communications Company, Inc
DT Newsletter
LA English
WC 85
TI ***Computer*** Speeds Genome Mapping
TX Argonne National Laboratory (9700 South Cass Avenue, Argonne, IL 60439; tel. 708-252-5584) has developed ***computer*** -automated technology for speeding the process of sequencing DNA, a technology

called "genome ***sequencing*** by ***hybridization*** ." It is a high-volume process which medical, agricultural and other companies, including reagent suppliers, already express interest in.

L64 ANSWER 19 OF 61 USPATFULL
AN 94:15495 USPATFULL
TI Parallel sequential reactor
IN Church, George M., Brookline, MA, United States
Kieffer-Higgins, Stephen G., Dorchester, MA, United States
PA The President and Fellows of Harvard College, Cambridge, MA, United States (U.S. corporation)
PI US 5288468 940222
AI US 92-970650 921030 (7)
RLI Continuation of Ser. No. US 91-705308, filed on 24 May 1991, now

L64 ANSWER 20 OF 61 CAPLUS COPYRIGHT 1996 ACS DUPLICATE 6
AN 1994:290915 CAPLUS
DN 120:290915
TI DNA ***sequencing*** by ***hybridization*** of oligonucleotide matrix. Calculation of continuous stacking hybridization efficiency
AU Lysov, Yu. P.; Chernyi, A. A.; Balaeff, A. A.; Keattie, K. L.; Mirzabekov, A. D.
CS W.A. Engelhardt Inst. Mol. Biol., Russian Acad. Sci., Moscow, 117984, Russia
SO J. Biomol. Struct. Dyn. (1994), 11(4), 797-812
CODEN: JBSDD6; ISSN: 0739-1102
DT Journal
LA English
AB In this paper the authors consider the efficiency of addnl. rounds of "continuous stacking" hybridization in DNA sequence reconstruction by hybridization with oligonucleotide matrix (SHOM). After the initial hybridization of target DNA with the matrix of oligonucleotides of fixed length L some addnl. hybridizations should be carried out in the presence of fluorescently labeled oligonucleotides of another length l. These addnl. oligonucleotides can hybridize in tandem with matrix tuples (continuous stacking hybridization) thus forming an extended duplex with the target DNA strand. The addnl. data obtained allows resolns. of branching points arising in the reconstruction procedure. Multiple rounds of continuous stacking hybridization considerably increase the efficiency of the sequencing method, eventually approaching the power of (L + l)-matrix. The authors develop here an ***algorithm*** that allows the authors to minimize the no. of addnl. hybridization steps, by assembling sets of l-tuples to be added together in each round of continuous stacking hybridization. For SHOM using a matrix of octanucleotides, continuous stacking hybridization with pentanucleotides increases the length of unambiguously sequenced DNA from 200 to several thousands of base pairs.
TI DNA ***sequencing*** by ***hybridization*** of oligonucleotide matrix. Calculation of continuous stacking hybridization efficiency
AB . . . increase the efficiency of the sequencing method, eventually approaching the power of (L + l)-matrix. The authors develop here an ***algorithm*** that allows the authors to minimize the no. of addnl. hybridization steps, by assembling sets of l-tuples to be added. . .
IT ***Algorithm***
(for DNA ***sequencing*** by ***hybridization*** of oligonucleotide matrix, continuous stacking hybridization in relation to)
IT Nucleotides, polymers
RL: BIOL (Biological study)
(oligo-, deoxyribo-, DNA ***sequencing*** by ***hybridization*** to matrixes, calcn. of continuous stacking

L64 ANSWER 27 OF 61 BIOTECHDS COPYRIGHT 1996 DERWENT INFORMATION LTD
AN 95-04430 BIOTECHDS
TI DNA ***sequencing*** by ***hybridization*** to oligonucleotides;
automated hybridization for potential use in mutation analysis, sequence comparison and mapping (conference abstract)
AU Mirzabekov A D
CS Inst.Mol.Biol.Moscow; Russian-Acad.Sci.
LO Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 117984, Russia.
SO Tools Genome Mapp.; (1994) 19
CODEN: 9999S
Tools for Genome Mapping, Paris, France, 16-18 January, 1994.
DT Journal
LA English
AB A DNA sequencing technique based on DNA hybridization with an oligonucleotide matrix (SHOM) was developed. The sequencing microchips consisted of a glass plate covered with 20 um-thick polyacrylamide gel squares (30-100 um square) each containing individual immobilized oligonucleotides (ODNs). The apparent thermostability of DNA duplexes with gel-immobilized ODNs increased with an increase in the ODN concentration and gel thickness. This allowed the concentration of immobilized ODNs to be adjusted in such a way as to equalize the thermostabilities of A-T- and G-C-rich duplexes. A prototype automatic sequenator was constructed consisting of a microchip, a fluorescent microscope, a CCD-camera, a thermostated plate and a ***computer*** complete with special software for image analysis. ***Computer*** simulations demonstrated that SHOM on octanucleotide matrices combined with continuous stacking hybridization could sequence 3,000-5,000 nucleotide long DNA. Chips for genetic mutation analysis, sequence comparison and DNA mapping are to be produced. (0 ref)
TI DNA ***sequencing*** by ***hybridization*** to oligonucleotides;
automated hybridization for potential use in mutation analysis, sequence comparison and mapping (conference abstract)
AB. . . A prototype automatic sequenator was constructed consisting of a microchip, a fluorescent microscope, a CCD-camera, a thermostated plate and a ***computer*** complete with special software for image analysis. ***Computer*** simulations demonstrated that SHOM on octanucleotide matrices combined with continuous stacking hybridization could sequence 3,000-5,000 nucleotide long DNA. Chips for. . .

```
L166            0 FILE HEALSAFE
L167            0 FILE PATDPA
L168            0 FILE PATOSDE
L169            0 FILE CEN
L170            0 FILE IPA
L171            0 FILE APIPAT
L172            0 FILE PAPERCHEM2
     TOTAL FOR ALL FILES
L173           33 S L87 AND (OLIGONUCLEOTIDE# OR NUCLEIC ACID# OR DNA OR PO
L174           23 DUPLICATE REMOVE L173 (10 DUPLICATES REMOVED)

=>
```

L64 ANSWER 2 OF 61 USPATFULL
AN 96:14707 USPATFULL
TI Method of determining an ordered sequence of subfragments of a nucleic acid fragment by hybridization of oligonucleotide probes
IN Drmanac, Radoje T., Beograd, Yugoslavia
Crkvenjakov, Radomir B., Beograd, Yugoslavia
PA Hyseq, Inc., Sunnyvale, CA, United States (U.S. corporation)
PI US 5492806 960220
AI US 93-45912 930412 (8)
DCD 20100413
RLI Continuation of Ser. No. US 91-723712, filed on 18 Jun 1991, now patented, Pat. No. US 5202231, issued on 13 Apr 1993 which is a continuation of Ser. No. US 88-175088, filed on 30 Mar 1988
PRAI YU 87-570 870401
DT Utility
EXNAM Primary Examiner: Zitomer, Stephanie W.
LREP Marshall, O'Toole, Gerstein, Murray & Borun
CLMN Number of Claims: 1
ECL Exemplary Claim: 1
DRWN No Drawings
LN.CNT 707
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB The sequence of a given nucleic acid fragment is read by the hybridization and assembly of positively hybridizing exactly complementary oligonucleotide probes through overlapping subfragments. By simultaneous hybridization of nucleic acid subfragments bound onto a filter, representing single-stranded phage vector with a cloned insert, with about 50,000 to 100,000 groups of probes, the main type of which is (A,T,C,G)(A,T,C,G)N8(A,T,C,G), information for ***computer*** determination of a sequence of DNA having the complexity of a mammalian genome are obtained in one step. To obtain a maximally completed sequence, three libraries cloned into the phage vector, M13, are used. The process can be easily and entirely robotized for factory reading of complex genomic fragments or DNA molecules.
AB . . . a cloned insert, with about 50,000 to 100,000 groups of probes, the main type of which is (A,T,C,G)(A,T,C,G)N8(A,T,C,G), information for ***computer*** determination of a sequence of DNA having the complexity of a mammalian genome are obtained in one step. To obtain. . .
SUMM . . . are laborious, with competent laboratories able to sequence approximately 100 bp per man per day. With the use of electronics (***computers*** and robots), sequencing can be accelerated by several orders of magnitude. The idea of sequencing the whole human genome has. . .
DETD The required synthesis of 4.times.10.sup.6 11-mers, is impracticable for ***sequencing*** by ***hybridization*** (SBH). However, it is unsuitable to omit a significant number of ONPs (more than 25%), because it leads to gaps. . .
DETD . . . generally speaking, ONSs repeated in tandems and having the length of one or more bp (AAAAAAA...TCTCTCTC...TGATGATG...) represent a problem in ***sequencing*** by ***hybridization*** . The above mentioned probes cannot determine length of repetitive sequences that are longer than the common part of a ONP. . .
DETD The ordering of SFs is performed by ***computer*** detection

L64 ANSWER 8 OF 61 CAPLUS COPYRIGHT 1996 ACS DUPLICATE 4
AN 1995:817552 CAPLUS
DN 123:331091
TI A new ***algorithm*** for DNA sequence assembly
AU Idury, Ramana M.; Waterman, Michael S.
CS Dep. Maths. Mol. Biol., Univ. Southern California, Los Angeles, CA, 90089-1113, USA
SO J. Comput. Biol. (1995), 2(2), 291-306
CODEN: JCOBEM; ISSN: 1066-5277
DT Journal
LA English
AB Since the advent of rapid DNA sequencing methods in 1976, scientist have had the problem of inferring DNA sequences from sequenced fragments. Shotgun sequencing is a well-established biol. and computational method used in practice. Many conventional ***algorithms*** for shotgun sequencing are based on the notion of pairwise fragment overlap. While shotgun sequencing infers a DNA sequence given the sequences of overlapping fragments, a recent and complementary method, called ***sequencing*** by ***hybridization*** (SBH), infers a DNA sequence given the set of oligomers that represents all subwords of some fixed length, k. In this paper, the authors propose a new ***computer*** ***algorithm*** for DNA sequence assembly that combines in a novel way the techniques of both shotgun and SBH methods. Based on our preliminary investigations, the ***algorithm*** promises to be very fast and practical for DNA sequence assembly.
TI A new ***algorithm*** for DNA sequence assembly
AB . . . inferring DNA sequences from sequenced fragments. Shotgun sequencing is a well-established biol. and computational method used in practice. Many conventional ***algorithms*** for shotgun sequencing are based on the notion of pairwise fragment overlap. While shotgun sequencing infers a DNA sequence given the sequences of overlapping fragments, a recent and complementary method, called ***sequencing*** by ***hybridization*** (SBH), infers a DNA sequence given the set of oligomers that represents all subwords of some fixed length, k. In this paper, the authors propose a new ***computer*** ***algorithm*** for DNA sequence assembly that combines in a novel way the techniques of both shotgun and SBH methods. Based on our preliminary investigations, the ***algorithm*** promises to be very fast and practical for DNA sequence assembly.
ST DNA sequence detn ***algorithm*** hybridization method
IT ***Algorithm***
Deoxyribonucleic acid sequence determination
(new ***algorithm*** for DNA sequence assembly involving ***sequencing*** by ***hybridization*** (SBH))
IT Nucleic acid hybridization
(DNA-DNA, ***sequencing*** by ***hybridization*** (SBH); new ***algorithm*** for DNA sequence assembly involving ***sequencing*** by ***hybridization*** (SBH))

L64 ANSWER 9 OF 61 CAPLUS COPYRIGHT 1996 ACS DUPLICATE 5
AN 1995:582209 CAPLUS
DN 123:26547
TI Clone clustering by hybridization
AU Milosavljevic, Aleksandar; Strezoska, Zaklina; Zeremski, Marija; Grujic, Danica; Paunesku, Tatjana; Crkvenjakov, Radomir
CS Genome Structure Group, Argonne National Lab., Argonne, IL, 60439-4833, USA
SO Genomics (1995), 27(1), 83-9
CODEN: GNMCEP; ISSN: 0888-7543
DT Journal
LA English
AB DNA ***sequencing*** by ***hybridization*** (SBH) Format 1 technique is based on expts. in which thousands of short oligomers are consecutively hybridized with dense arrays of clones. In this paper we present the description of a method for obtaining hybridization signatures for individual clones that guarantees reproducibility despite a wide range of variations is exptl. circumstances, a sensitive method for signature comparison at prespecified significance levels, and a clustering ***algorithm*** that correctly identifies clusters of significantly similar signatures. The methods and the ***algorithm*** have been verified exptl. on a control set of 422 signatures that originate from 9 distinct clones of known sequence. Expts. indicate that only 30 to 50 oligomer probes suffice for correct clustering. This information about the identity of clones can be used to guide both genomic and cDNA sequencing by SBH or by std. gel-based methods.
AB DNA ***sequencing*** by ***hybridization*** (SBH) Format 1 technique is based on expts. in which thousands of short oligomers are consecutively hybridized with dense arrays. . . wide range of variations is exptl. circumstances, a sensitive method for signature comparison at prespecified significance levels, and a clustering ***algorithm*** that correctly identifies clusters of significantly similar signatures. The methods and the ***algorithm*** have been verified exptl. on a control set of 422 signatures that originate from 9 distinct clones of known sequence.. . .
IT ***Algorithm***
Genetic methods
Nucleic acid hybridization
(clone clustering by hybridization using a clustering ***algorithm*** that correctly identifies clusters of significantly similar signatures)

N 95:435002 PROMT

TI Hyseq Licenses High-Speed Gene-Sequencing Technology
To license its ***Sequencing*** -By- ***Hybridization*** technology for the development of diagnostic tests

SO Genetic Engineering News, (15 Jan 1995) pp. 8.
ISSN: 0270-6377.

AB Hyseq of Sunnyvale, CA is to license its ***Sequencing*** -By- ***Hybridization*** (SBH) patented technology in its Format-1 variation. The license will be available to companies who are developing diagnostic tests for genetic and infectious diseases. SBH allows speedy detection of gene mutations, the program being able to provide information on drug resistance as well as mutational detection in regions of genes. Image analysis ***computer*** software is used to analyze results. The program can be adapted for use with specific disease programs including cancer and cystic fibrosis.

TI Hyseq Licenses High-Speed Gene-Sequencing Technology
To license its ***Sequencing*** -By- ***Hybridization*** technology for the development of diagnostic tests
Hyseq of Sunnyvale, CA is to license its ***Sequencing*** -By- ***Hybridization*** (SBH) patented technology in its Format-1 variation. The license will be available to companies who are developing diagnostic tests for. . . program being able to provide information on drug resistance as well as mutational detection in regions of genes. Image analysis ***computer*** software is used to analyze results. The program can be adapted for use with specific disease programs including cancer and. . .

L64 ANSWER 34 OF 61 DISSABS COPYRIGHT 1996 UMI
AN 94:28331 DISSABS Order Number: AAR9417770
TI COMBINATORIAL ***ALGORITHMS*** FOR COMPUTATIONAL BIOLOGY (DNA SEQUENCING, RESTRICTION MAPPING)
AU SUNDARAM, GOPALAKRISHNAN [PH.D.]; SKIENA, STEVEN S. [adviser]
CS STATE UNIVERSITY OF NEW YORK AT STONY BROOK (0771)
SO Dissertation Abstracts International, (1993) Vol. 55, No. 2B, p. 486. Order No.: AAR9417770. 147 pages.
DT Dissertation
FS DAI
LA English
AB In this dissertation, we address some of the algorithmic and computational issues involved in computational biology problems. In particular, we give new ***algorithms*** for three different problems, viz., restriction site mapping, reconstructing DNA strings from ***sequencing*** by ***hybridization*** (SBH) data, and subgraph detection ***algorithms*** .

We propose new, practical ***algorithms*** which resolve the noisy experimental data from DNA partial digests, including the possibility of missing fragments. We analyze the performance of these ***algorithms*** under a reasonable probability distribution, establishing a relative error limit of r = Θ(1/$n\sp2$) beyond which our technique becomes infeasible. Through simulations, we establish that our technique is robust enough to reconstruct data with relative errors of up to 7.0% in the measured fragment lengths for typical problems.

Sequencing by ***hybridization*** is a new approach to DNA sequencing in which we sequence the DNA by asking all substring queries of length m, using a sequencing chip C(m). We develop a theory of interactive ***sequencing*** by ***hybridization*** , based on reconstructing strings from substrings. We provide tight bounds on the complexity of reconstructing unknown strings from substring queries. Specifically, we show that $\Theta({n}\alpha)$ substring queries are necessary and sufficient, where α is the alphabet size. We also demonstrate that subsequence queries are more efficient by showing O(n lg α) queries are necessary and sufficient. We give ***algorithms*** to build decision trees which are within a constant factor of an optimal decision tree with the constant depending upon α.

Finally, we consider the problem of subgraph isomorphism which has applications in fragment assembly and other DNA sequencing ***algorithms*** . We present general ***algorithms*** for fixed-subgraph isomorphism which improve or unify previous results. In particular, we present an $O(n\sp{f}m)$ ***algorithm*** for recognizing a fixed subgraph H with flower number f within a graph G with n vertices and m edges. (Abstract shortened by UMI.)
TI COMBINATORIAL ***ALGORITHMS*** FOR COMPUTATIONAL BIOLOGY (DNA SEQUENCING, RESTRICTION MAPPING)
AB . . . dissertation, we address some of the algorithmic and computational issues involved in computational biology problems. In particular, we give new ***algorithms*** for three different problems, viz., restriction site mapping, reconstructing DNA strings from ***sequencing*** by ***hybridization*** (SBH) data, and subgraph detection ***algorithms*** .

We propose new, practical ***algorithms*** which resolve

the noisy experimental data from DNA partial digests, including the possibility of missing fragments. We analyze the performance of these ***algorithms*** under a reasonable probability distribution, establishing a relative error limit of r = Θ(1/$n\sp2$) beyond which our technique becomes infeasible.. . . robust enough to reconstruct data with relative errors of up to 7.0% in the measured fragment lengths for typical problems.

Sequencing by ***hybridization*** is a new approach to DNA sequencing in which we sequence the DNA by asking all substring queries of length m, using a sequencing chip C(m). We develop a theory of interactive ***sequencing*** by ***hybridization*** , based on reconstructing strings from substrings. We provide tight bounds on the complexity of reconstructing unknown strings from substring queries.. . . also demonstrate that subsequence queries are more efficient by showing O(n lg α) queries are necessary and sufficient. We give ***algorithms*** to build decision trees which are within a constant factor of an optimal decision tree with the constant depending upon α.

Finally, we consider the problem of subgraph isomorphism which has applications in fragment assembly and other DNA sequencing ***algorithms*** . We present general ***algorithms*** for fixed-subgraph isomorphism which improve or unify previous results. In particular, we present an $O(n\sp{f}m)$ ***algorithm*** for recognizing a fixed subgraph H with flower number f within a graph G with n vertices and m edges.. . .

L64 ANSWER 35 OF 61 NTIS COPYRIGHT 1996 NTIS
AN 93(24):1154 NTIS Order Number : DE93015556/XAD
TI Discovering sequence similarity by the algorithmic significance method.
AU Milosavljevic, A.
CS Argonne National Lab., IL
Sponsor : Department of Energy, Washington, DC
NC Contract : W-31109-ENG-38; FG03-91ER61152
NR DE93015556/XAD; ANL/BIM/CP-78918; CONF-930745-2
11 p. NTIS Prices : PC A03/MF A01
Notes : International conference in intelligent systems for molecular biology (1st), Washington, DC (United States), 7-9 Jul 1993. Sponsored by Department of Energy, Washington, DC.
PD Feb 1993
LA English CY United States
OS ERA9350
AB The minimal-length encoding approach is applied to define concept of sequence similarity. A sequence is defined to be similar to another sequence or to a set of keywords if it can be encoded in a small number of bits by taking advantage of common subwords. Minimal-length encoding of a sequence is computed in linear time, using a data compression algorithm that is based on a dynamic programming strategy and the directed acyclic word graph data structure. No assumptions about common word (''k-tuple'') length are made in advance, and common words of any length are considered. The newly proposed algorithmic significance method provides an exact upper bound on the probability that sequence similarity has occurred by chance, thus eliminating the need for any arbitrary choice of similarity thresholds. Preliminary experiments indicate that a small number of keywords can positively identify a DNA sequence, which is extremely relevant in the context of partial sequencing by hybridization.

TI Sequencing of megabase plus DNA by hybridization: theory of the method

AU Drmanac, Radoje; Labat, Ivan; Brukner, Ivan; Crkvenjakov, Radomir
CS Genet. Eng. Cent., Belgrade, 11000, Yugoslavia

SO Genomics (1989), 4(2), 114-28
CODEN: GNMCEP; ISSN: 0888-7543
DT Journal
LA English

AB A theory for mismatch-free hybridization of oligonucleotides contg. from 11 to 20 monomers to unknown target DNA for sequencing of large genomes is presented. Probability calcns. and, in part, ***computer*** simulations were used to est. the types and nos. of oligonucleotides that would have to be synthesized in order to sequence a megabase plus segment of DNA. It was estd. that 95,000 specific mixes of 11-mers, mainly of the 5' (A,T,C,G)(A,T,C,G)N8(A,T,C,G)3' type, hybridized consecutively to dot blots of cloned genomic DNA fragments would provide primary data for the sequence assembly. An optimal mixt. of representative libraries in M13 vector, having inserts of (1) 7 kb, (2) 0.5 kb genomic fragments randomly ligated in up to 10-kb inserts, and (3) tandem jumping fragments 100 kb apart in the genome, would be needed. To sequence each million base pairs of DNA, hybridization data from about 2100 sep. hybridization sample dots would be required. Inevitable gaps and uncertainties in alignment of sequenced fragments were considered and minimized by choice of libraries and no. of subclones used for hybridization. Because it is based on simpler biochem. procedures, this method is inherently easier to automate than existing sequencing methods. The sequence can be derived from simple primary data only by extensive computing. Phased exptl. tests and ***computer*** simulations are required for accurate ests. in terms of cost and speed of sequencing by the new approach. Nevertheless, ***sequencing*** by ***hybridization*** has apparent advantages over existing methods because of the inherent redundancy and parallelism in its data gathering.

AB . . . 11 to 20 monomers to unknown target DNA for sequencing of large genomes is presented. Probability calcns. and, in part, ***computer*** simulations were used to est. the types and nos. of oligonucleotides that would have to be synthesized in order to. . . existing sequencing methods. The sequence can be derived from simple primary data only by extensive computing. Phased exptl. tests and ***computer*** simulations are required for accurate ests. in terms of cost and speed of sequencing by the new approach. Nevertheless, ***sequencing*** by ***hybridization*** has apparent advantages over existing methods because of the inherent redundancy and parallelism in its data gathering.

=>

L64 ANSWER 38 OF 42 CAPLUS COPYRIGHT 1996 ACS DUPLICATE 5
AN 1992:585718 CAPLUS
DN 117:185718
TI An oligonucleotide matrix hybridization approach to DNA sequencing
AU Khorlin, A. A.; Khrapko, K. R.; Ivanov, I. B.; Lysov, Yu. P.; Ershov, G. K.; Vasilenko, S. K.; Florent'ev, V. L.; Mirzabekov, A. D.
CS V. A. Engel'gardt Inst. Mol. Biol., Moscow, 117984, Russia
SO Nucleic Acids Symp. Ser. (1991), 24(Synth. Oligonucleotides: Probl. Front. Pract. Appl.), 191-2
CODEN: NACSD8; ISSN: 0261-3166
DT Journal
LA English
AB A new approach to DNA ***sequencing*** by ***hybridization*** with oligonucleotide matrix (SHOM) which could also be applied for DNA mapping and fingerprinting, mutant diagnostics, etc., has been tested in model expts. A dot matrix was prepd. which contained 9 overlapping octanucleotides (8-mers) complementary to a common 17-mer. Each of the 8-mers (5 pmoles) contg. 3-N-methyluridine at the 3'-terminus was immobilized after periodate oxidn. as individual dot (1 mm in diam.) in thin layer (30 .mu.m-thick) of hydrazine-activated polyacrylamide gel fixed on a glass plate. The matrix was sep. hybridized with 32P-labeled 17-mer and three other 17-mers differing from the first one by a single base change. This set of fragments and octanucleotides can form perfect duplexes as well as defective ones, each contg. one out of five base pair mismatches (A:G, A:C, G:T, C:T and C:C), differently positioned relatively to the duplex ends. There always exists a certain stage of thermal dissocn. at which the hybridization signals ***ratio*** of perfect to imperfect duplex is sufficiently high (at least 10 times). In that way dissocn. curves comparison allow one to distinguish perfect duplexes from mismatched ones and thus to detect single base changes in DNA. SHOM with full octanucleotide matrix can be applied for sequencing much longer DNA as well for localization and identification of DNA changes in homologous sequences and gene mutations. Miniaturized matrixes or sequencing chips were designed where oligonucleotides were immobilized within 100 .times. 100 .mu. dots disposed at 100 .mu. intervals. Hybridization of fluorescently labeled DNA fragments with micro chips may simplify sequencing and ensure sensitivity of at least 10 amol per dot.
AB A new approach to DNA ***sequencing*** by ***hybridization*** with oligonucleotide matrix (SHOM) which could also be applied for DNA mapping and fingerprinting, mutant diagnostics, etc., has been tested. . . positioned relatively to the duplex ends. There always exists a certain stage of thermal dissocn. at which the hybridization signals ***ratio*** of perfect to imperfect duplex is sufficiently high (at least 10 times). In that way dissocn. curves comparison allow one. . .

L64 ANSWER 37 OF 42 BIOTECHDS COPYRIGHT 1996 DERWENT INFORMATION LTD
AN 94-00636 BIOTECHDS
TI Towards genome DNA sequencing chip based on oligonucleotide hybridization;
miniaturization for use in the human genome project (conference paper)
AU Drmanac R; Strezoska Z; Labat I; Radosavljevic D; Paunesku T; Crkvenjakov R
CS Inst.Mol.Genet.+Genet.Eng.Belgrade
LO Institute for Molecular Genetics and Genetic Engineering, Belgrade, Yugoslavia.
SO Modelling Computer Methods Mol.Biol.Genet.; (1991) 242-43
CODEN: 9999U
DT Journal
LA English
AB Miniaturization of DNA ***sequencing*** by ***hybridization*** (which involves reconstruction of a sequence from complete content of short 8-mer oligonucleotides) was carried out, to improve the speed of genome sequencing. The proposed DNA sequencing chip is a microhybridization surface, on which all necessary oligos of known formula occur at physically defined positions. The chip may have 100,000-1,000 million oligos. After hybridization of a chip with a genomic fragment, a pattern of positive dots is fed into a computer by microscopic imaging. The computer sorts the data from many hybridizations into a genomic sequence. The basic step in chip manufacture is mixed combinatorial synthesis of oligos on beads marked with specific combinations of marker oligos. Conditions for reliable hybridization of 8-mers have been determined. The ***ratio*** of true to false signals is 3-30, depending on an 8-mer sequence in model phage M13 clones with a 1-kb insert, indicating the reaction is usable for sequencing. An algorithm has been formulated for generation of sequences between 2 points of ambiguity from hybridization data from a gene bank. (1 ref)
AB Miniaturization of DNA ***sequencing*** by ***hybridization*** (which involves reconstruction of a sequence from complete content of short 8-mer oligonucleotides) was carried out, to improve the speed. . . oligos on beads marked with specific combinations of marker oligos. Conditions for reliable hybridization of 8-mers have been determined. The ***ratio*** of true to false signals is 3-30, depending on an 8-mer sequence in model phage M13 clones with a 1-kb. . .

L64 ANSWER 28 OF 42 CAPLUS COPYRIGHT 1996 ACS
AN 1994:571510 CAPLUS
DN 121:171510
TI Simulations of ordering and sequence reconstruction of random DNA clones hybridized with a small number of oligomeric probes
AU Labat, Ivan; Drmanac, Radoje
CS Biol. Med. Res. Div., Argonne Natl. Lab., Argonne, IL, 60439, USA
SO Int. Conf. Bioinformatics, Supercomput. Complex Genome Anal., Proc. Conf., 2nd (1993), Meeting Date 1992, 555-65. Editor(s): Lim, Hwa A. Publisher: World Sci. Publ., Singapore, Singapore.
CODEN: 59QBA7
DT Conference
LA English
AB The ***sequencing*** by ***hybridization*** (SBH) method has been developed for assaying millions of 0.5-2-kb-long clones. This opens up an efficient way for defining the order of short clones and creating a phys. map at 100-bp resoln. Moreover, complete sequences can be obtained using a modest no. (.apprx.3000) of probes if hybridization and gel sequence data from overlapped or similar sequences are used. In ***light*** of these possibilities, various heuristic algorithms have been developed and tested in simulation expts. This approach can influence the interpretation of the intuitively obvious term, known sequence.
AB The ***sequencing*** by ***hybridization*** (SBH) method has been developed for assaying millions of 0.5-2-kb-long clones. This opens up an efficient way for defining the. . . a modest no. (.apprx.3000) of probes if hybridization and gel sequence data from overlapped or similar sequences are used. In ***light*** of these possibilities, various heuristic algorithms have been developed and tested in simulation expts. This approach can influence the interpretation. . .

L64 ANSWER 25 OF 42 TOXLIT
AN 94:115986 TOXLIT
DN CA-121-171510N
TI Simulations of ordering and sequence reconstruction of random DNA clones hybridized with a small number of oligomeric probes.
AU Labat I; Drmanac R
CS Biol. Med. Res. Div., Argonne Natl. Lab., Argonne
SO Int. Conf. Bioinformatics, Supercomput. Complex Genome Anal., Proc. Conf., 2nd, (1993). pp. 555-65.
CODEN: 59QBA.
CY United States
DT Book; (MONOGRAPH)
FS CA
LA English
OS CA 121:171510
EM 9411
AB The ***sequencing*** by ***hybridization*** (SBH) method has been developed for assaying millions of 0.5-2-kb-long clones. This opens up an efficient way for defining the order of short clones and creating a phys. map at 100-bp resoln. Moreover, complete sequences can be obtained using a modest no. (.apprx.3000) of probes if hybridization and gel sequence data from overlapped or similar sequences are used. In ***light*** of these possibilities, various heuristic algorithms have been developed and tested in simulation expts. This approach can influence the interpretation of the intuitively obvious term, known sequence.
AB The ***sequencing*** by ***hybridization*** (SBH) method has been developed for assaying millions of 0.5-2-kb-long clones. This opens up an efficient way for defining the. . . a modest no. (.apprx.3000) of probes if hybridization and gel sequence data from overlapped or similar sequences are used. In ***light*** of these possibilities, various heuristic algorithms have been developed and tested in simulation expts. This approach can influence the interpretation. . .

L64 ANSWER 24 OF 42 USPATFULL
AN 93:29111 USPATFULL
TI Method of sequencing of genomes by hybridization of oligonucleotide probes
IN Drmanac, Radoje T., Zvecanska 46, Beograd, Yugoslavia 11000
Crkvenjakov, Radomir B., Bulevar JNA 118, Beograd, Yugoslavia 11000
PI US 5202231 930413
AI US 91-723712 910618 (7)
RLI Continuation of Ser. No. US 88-175088, filed on 30 Mar 1988, now abandoned
PRAI YU 87-570 870401
DT Utility
EXNAM Primary Examiner: Moskowitz, Margaret; Assistant Examiner: Zitomer, Stephanie W.
LREP Marshall, O'Toole, Gerstein, Murray & Bicknell
CLMN Number of Claims: 4
ECL Exemplary Claim: 1
DRWN No Drawings
LN.CNT 673
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB The conditions under which oligonucleotides hybridize only with entirely homologous sequences are recognized. The sequence of a given DNA fragment is read by the hybridization and assembly of positively hybridizing probes through overlapping portions. By simultaneous hybridization of DNA molecules applied as dots and bound onto a filter, representing single-stranded phage vector with the cloned insert, with about 50,000 to 100,000 groups of probes, the main type of which is (A,T,C,G)(A,T,C,G)N8(A,T,C,G), information for computer determination of a sequence of DNA having the complexity of a mammalian genome are obtained in one step. To obtain a maximally completed sequence, three libraries cloned into the phage vector, M13, bacteriophage are used: with the 0.5 kb and 7 kbp insert consisting of two sequences, with the average distance in genomic DNA of 100 kbp. For a million bp of genomic DNA, 25,000 subclones of the 0.5 kbp are required as well as 700 subclones 7 kb long and 170 jumping subclones. Subclones of 0.5 kb are applied on a filter in groups of 20 each, so that the total number of samples is 2,120 per million bp. The process can be easily and entirely robotized for factory reading of complex genomic fragments or DNA molecules.
DETD The required synthesis of 4.times.10.sup.6 11-mers, is impracticable for ***sequencing*** by ***hybridization*** (SBH). However, it is unsuitable to omit a significant number of ONPs (more than 25%), because it leads to gaps. . .
DETD . . . or more bp (AAAAAAA . . . TCTCTCTC . . . TGATGATG . . .) represent a problem in ***sequencing*** by ***hybridization*** . The above mentioned probes cannot determine length of repetitive sequences that are longer than the common part of a ONP. . .
DETD . . . random order. Such ONPs will give unreliable information or cannot be used at all if they give signals of approximate ***strength*** to signals of all colonies.

L64 ANSWER 21 OF 42 CABA COPYRIGHT 1996 CABI
AN 94:71047 CABA
DN 940104049
TI Genosensors: microfabricated devices for automated DNA sequence analysis
AU Eggers, M. D.; Hogan, M. E.; Reich, R. K.; Lamture, J. B.; Beattie, K. L.; Hollis, M. A.; Ehrlich, D. J.; Kosicki, B. B.; Shumaker, J. M.; Varma, R. S.; Burke, B. E.; Murphy, A.; Rathman, D. D.
CS Houston Advanced Research Center, The Woodlands, TX 77381, USA.
SO (1993) Vol. 1891, pp. 113-126. 19 ref.
Publisher: Society of Photo-Optical Instrumentation Engineers (SPIE). Bellingham
Meeting Info.: Proceedings of SPIE.
CY United States
DT Miscellaneous
LA English
AB A new technology is described for developing low cost, high throughput DNA ***sequencing*** by ***hybridization*** . It uses bioelectronic sensor devices to detect hybridization across a DNA probe array. Genosensors contain numerous micro-sized electronic test fixtures assembled in a 2-dimensional array. These electronically addressable fixtures contain immobilized synthetic DNA probes which hybridize to target DNA samples applied to the array. Some results with a charge-coupled device are described, in which hybridization with labelled target DNA to an array of probes immobilized above the ***pixels*** was carried out. It was shown that such coupling provides a 10-fold improvement in sensitivity over conventional lens-based detection systems. The preliminary results were obtained with low-density arrays, but mathematical analysis supports the development of higher density arrays suitable for sequencing and diagnostic applications.
AB A new technology is described for developing low cost, high throughput DNA ***sequencing*** by ***hybridization*** . It uses bioelectronic sensor devices to detect hybridization across a DNA probe array. Genosensors contain numerous micro-sized electronic test fixtures. . . a charge-coupled device are described, in which hybridization with labelled target DNA to an array of probes immobilized above the ***pixels*** was carried out. It was shown that such coupling provides a 10-fold improvement in sensitivity over conventional lens-based detection systems.. . .

=> s (sequencing by hybridization)

L2	101	FILE	CAPLUS
L3	82	FILE	BIOSIS
L4	75	FILE	TOXLIT
L5	74	FILE	MEDLINE
L6	56	FILE	SCISEARCH
L7	47	FILE	EMBASE
L8	44	FILE	BIOTECHDS
L9	39	FILE	LIFESCI
L10	21	FILE	NLDB
L11	19	FILE	CABA
L12	19	FILE	PROMT
L13	18	FILE	USPATFULL
L14	13	FILE	CANCERLIT
L15	12	FILE	BIOBUSINESS
L16	9	FILE	DISSABS
L17	8	FILE	NTIS
L18	7	FILE	TOXLINE
L19	7	FILE	PNI
L20	7	FILE	INPADOC
L21	6	FILE	CEABA
L22	5	FILE	EMBAL
L23	4	FILE	CEN
L24	4	FILE	CIN
L25	4	FILE	CJACS
L26	4	FILE	CONFSCI
L27	4	FILE	JICST-EPLUS
L28	2	FILE	AQUASCI
L29	1	FILE	FSTA
L30	1	FILE	IFIPAT
L31	1	FILE	OCEAN

TOTAL FOR ALL FILES
L32 694 (SEQUENCING BY HYBRIDIZATION)

=> s l32 and (intensit### or ratio# or strength or pixel# or light)

L33	5	FILE	CAPLUS
L34	4	FILE	BIOSIS
L35	5	FILE	TOXLIT
L36	3	FILE	MEDLINE
L37	3	FILE	SCISEARCH
L38	3	FILE	EMBASE
L39	5	FILE	BIOTECHDS
L40	1	FILE	LIFESCI
L41	1	FILE	NLDB
L42	1	FILE	CABA
L43	0	FILE	PROMT
L44	15	FILE	USPATFULL
L45	0	FILE	CANCERLIT
L46	0	FILE	BIOBUSINESS
L47	0	FILE	DISSABS
L48	1	FILE	NTIS
L49	0	FILE	TOXLINE
L50	0	FILE	PNI
L51	0	FILE	INPADOC

```
L52             2 FILE CEABA
L53             0 FILE EMBAL
L54             3 FILE CEN
L55             0 FILE CIN
L56             4 FILE CJACS
L57             0 FILE CONFSCI
L58             0 FILE JICST-EPLUS
L59             0 FILE AQUASCI
L60             0 FILE FSTA
L61             1 FILE IFIPAT
L62             0 FILE OCEAN

TOTAL FOR ALL FILES
L63            57 L32 AND (INTENSIT### OR RATIO# OR STRENGTH OR PIXEL# OR LI
                  GHT)

=> duplicate remove l63
```

L189 ANSWER 419 OF 437 USPATFULL
AN 81:5454 USPATFULL
TI Method and apparatus for physiologic facsimile imaging of biologic targets based on complex permittivity measurements using remote microwave interrogation
IN Larsen, Lawrence E., Silver Spring, MD, United States
Jacobi, John H., Bowie, MD, United States
PA The United States of America as represented by the Secretary of the Army, Washington, DC, United States (U.S. government)
PI US 4247815 810127
AI US 79-41374 790522 (6)
RLI Continuation-in-part of Ser. No. US 77-891256, filed on 14 Oct 1977, now patented, Pat. No. US 4162500 which is a continuation-in-part of Ser. No. US 77-842137, filed on 14 Oct 1977, now patented, Pat. No. US 4135131
DT Utility
EXNAM Primary Examiner: Krawczewicz, Stanley T.
LREP Gapcynski, William G.; Winters, Sherman D.; Bellamy, Werten F. W.
CLMN Number of Claims: 16
ECL Exemplary Claim: 1
DRWN 10 Drawing Figure(s); 8 Drawing Page(s)
LN.CNT 856
AB A physiologic facsimile image of a biological target is obtained by: scang the target by transmitting a microwave signal through the target and measuring at least one of the amplitude and phase components of the complex microwave power transmission coefficient at each one of a plurality of sample locations which are spaced so as to define a two-dimensional ***array*** and such that a set of digital data for each of the measured components is obtained, and for at least one of the sets of data; sorting the set of data into column order; magnifying data derived from the sorting step so as to enhance the resolution of the image; mapping data derived from the magnifying step into further data using a predetermined mapping function so as to enhance the contrast between selected portions of the image; and obtaining the set of control signals by filtering data derived from the mapping step using a band pass function which rejects spatial frequencies below a predetermined first frequency and/or rejects spatial frequencies above a predetermined second frequency so as to minimize, respectively, the effects of variations in the thickness of the target and/or spurious frequencies resulting from the magnifying step.

```
L32          694 S (SEQUENCING BY HYBRIDIZATION)
L33            5 FILE CAPLUS
L34            4 FILE BIOSIS
L35            5 FILE TOXLIT
L36            3 FILE MEDLINE
L37            3 FILE SCISEARCH
L38            3 FILE EMBASE
L39            5 FILE BIOTECHDS
L40            1 FILE LIFESCI
L41            1 FILE NLDB
L42            1 FILE CABA
L43            0 FILE PROMT
L44           15 FILE USPATFULL
L45            0 FILE CANCERLIT
L46            0 FILE BIOBUSINESS
L47            0 FILE DISSABS
L48            1 FILE NTIS
L49            0 FILE TOXLINE
L50            0 FILE PNI
L51            0 FILE INPADOC
L52            2 FILE CEABA
L53            0 FILE EMBAL
L54            3 FILE CEN
L55            0 FILE CIN
L56            4 FILE CJACS
L57            0 FILE CONFSCI
L58            0 FILE JICST-EPLUS
L59            0 FILE AQUASCI
L60            0 FILE FSTA
L61            1 FILE IFIPAT
L62            0 FILE OCEAN
     TOTAL FOR ALL FILES
L63           57 S L32 AND (INTENSIT### OR RATIO# OR STRENGTH OR PIXEL# OR
L64           42 DUPLICATE REMOVE L63 (15 DUPLICATES REMOVED)
L65           61 FILE CAPLUS
L66           15 FILE BIOSIS
L67           14 FILE TOXLIT
L68           11 FILE MEDLINE
L69           53 FILE SCISEARCH
L70           13 FILE EMBASE
L71            9 FILE BIOTECHDS
L72           10 FILE LIFESCI
L73           33 FILE NLDB
L74            7 FILE CABA
L75           90 FILE PROMT
L76         1638 FILE USPATFULL
L77            1 FILE CANCERLIT
L78            1 FILE BIOBUSINESS
L79           21 FILE DISSABS
L80           21 FILE NTIS
L81            2 FILE TOXLINE
L82            0 FILE PNI
L83            0 FILE INPADOC
L84            0 FILE CEABA
L85            1 FILE EMBAL
L86            4 FILE CEN
L87            1 FILE CIN
```

```
L88          152 FILE CJACS
L89            0 FILE CONFSCI
L90            5 FILE JICST-EPLUS
L91            2 FILE AQUASCI
L92            0 FILE FSTA
L93           66 FILE IFIPAT
L94            2 FILE OCEAN
    TOTAL FOR ALL FILES
L95         2233 S (MISMATCHE#) AND (ARRAY# OR SEQUENCING BY HYBRIDIZATION
L96            9 FILE CAPLUS
L97            3 FILE BIOSIS
L98            3 FILE TOXLIT
L99            2 FILE MEDLINE
L100          15 FILE SCISEARCH
L101           3 FILE EMBASE
L102           2 FILE BIOTECHDS
L103           2 FILE LIFESCI
L104          19 FILE NLDB
L105           2 FILE CABA
L106          48 FILE PROMT
L107        1149 FILE USPATFULL
L108           0 FILE CANCERLIT
L109           0 FILE BIOBUSINESS
L110          10 FILE DISSABS
L111           8 FILE NTIS
L112           0 FILE TOXLINE
L113           0 FILE PNI
L114           0 FILE INPADOC
L115           0 FILE CEABA
L116           0 FILE EMBAL
L117           3 FILE CEN
L118           0 FILE CIN
L119         129 FILE CJACS
L120           0 FILE CONFSCI
L121           2 FILE JICST-EPLUS
L122           1 FILE AQUASCI
L123           0 FILE FSTA
L124          15 FILE IFIPAT
L125           1 FILE OCEAN
     TOTAL FOR ALL FILES
L126        1426 S L95 AND (ALGORITHM# OR COMPUTER# OR INTENSIT### OR RATI
L127           7 FILE CAPLUS
L128           1 FILE BIOSIS
L129           1 FILE TOXLIT
L130           0 FILE MEDLINE
L131           8 FILE SCISEARCH
L132           1 FILE EMBASE
L133           0 FILE BIOTECHDS
L134           0 FILE LIFESCI
L135           1 FILE NLDB
L136           0 FILE CABA
L137           6 FILE PROMT
L138         821 FILE USPATFULL
L139           0 FILE CANCERLIT
L140           0 FILE BIOBUSINESS
L141           5 FILE DISSABS
L142           2 FILE NTIS
L143           0 FILE TOXLINE
```

```
L144            0 FILE PNI
L145            0 FILE INPADOC
L146            0 FILE CEABA
L147            0 FILE EMBAL
L148            2 FILE CEN
L149            0 FILE CIN
L150          113 FILE CJACS
L151            0 FILE CONFSCI
L152            1 FILE JICST-EPLUS
L153            0 FILE AQUASCI
L154            0 FILE FSTA
L155            9 FILE IFIPAT
L156            0 FILE OCEAN
      TOTAL FOR ALL FILES
L157          978 S L126 AND (RATIO# OR INTENSIT###)
L158            0 FILE CAPLUS
L159            0 FILE BIOSIS
L160            0 FILE TOXLIT
L161            0 FILE MEDLINE
L162            2 FILE SCISEARCH
L163            0 FILE EMBASE
L164            0 FILE BIOTECHDS
L165            0 FILE LIFESCI
L166            1 FILE NLDB
L167            0 FILE CABA
L168            0 FILE PROMT
L169          380 FILE USPATFULL
L170            0 FILE CANCERLIT
L171            0 FILE BIOBUSINESS
L172            2 FILE DISSABS
L173            0 FILE NTIS
L174            0 FILE TOXLINE
L175            0 FILE PNI
L176            0 FILE INPADOC
L177            0 FILE CEABA
L178            0 FILE EMBAL
L179            1 FILE CEN
L180            0 FILE CIN
L181           51 FILE CJACS
L182            0 FILE CONFSCI
L183            0 FILE JICST-EPLUS
L184            0 FILE AQUASCI
L185            0 FILE FSTA
L186            2 FILE IFIPAT
L187            0 FILE OCEAN
      TOTAL FOR ALL FILES
L188          439 S L157 AND (COMPUTER OR ALGORITHM)
L189          437 DUPLICATE REMOVE L188 (2 DUPLICATES REMOVED)
L190            0 FILE CAPLUS
L191            0 FILE BIOSIS
L192            0 FILE TOXLIT
L193            0 FILE MEDLINE
L194            0 FILE SCISEARCH
L195            0 FILE EMBASE
L196            0 FILE BIOTECHDS
L197            0 FILE LIFESCI
L198            0 FILE NLDB
L199            0 FILE CABA
```

```
L200            0 FILE PROMT
L201           21 FILE USPATFULL
L202            0 FILE CANCERLIT
L203            0 FILE BIOBUSINESS
L204            0 FILE DISSABS
L205            0 FILE NTIS
L206            0 FILE TOXLINE
L207            0 FILE PNI
L208            0 FILE INPADOC
L209            0 FILE CEABA
L210            0 FILE EMBAL
L211            1 FILE CEN
L212            0 FILE CIN
L213           30 FILE CJACS
L214            0 FILE CONFSCI
L215            0 FILE JICST-EPLUS
L216            0 FILE AQUASCI
L217            0 FILE FSTA
L218            0 FILE IFIPAT
L219            0 FILE OCEAN
     TOTAL FOR ALL FILES
L220           52 S L188 AND (OLIGONUCLEOTIDE# OR NUCLEIC ACID# OR DNA)
L221           52 DUPLICATE REMOVE L220 (0 DUPLICATES REMOVED)

=>
```

of pools containing subclones which overlap with the starting SSF. This detection is performed on the basis of the. . .

DETD . . . hybridization, the technological procedure is continued by reading of results of hybridization. Data are stored in the memory of the ***computer*** center. Data have binary characters (+,-) and are read from several sensitivity thresholds. Based on these, SFs are first formed,. . . it is followed by mutual ordering of SFs and subclones. At the end of processing of all the data, the ***computer*** center determines which subclones must be treated experimentally and what type of treatment should be applied in order to obtain. . .

DETD SBH is the method which minimizes experimental work at the expense of additional ***computer*** work. The only technological requirement is the sequence-specific hybridization of ONP. An incapability to use up to 6% of predicted. . .

DETD Solutions upon which this method is based enable one to obtain enough data in an entirely automated, ***computer*** -guided plant from data in the form of binary signals; computing generates the sequence of complex, cloned DNA fragments and/or molecules,. . .

L64 ANSWER 3 OF 61 CAPLUS COPYRIGHT 1996 ACS DUPLICATE 1
AN 1996:259328 CAPLUS
TI Positional ***sequencing*** by ***hybridization***
AU Hannenhalli, Sridhar; Feldman, William; Lewis, Herbert F.; Skiena, Steven S.; Pevzner, Pavel A.
CS Department Computer Science and Engineering, Pennsylvania State University, University Park, PA, 16802, USA
SO Comput. Appl. Biosci. (1996), 12(1), 19-24
CODEN: COABER; ISSN: 0266-7061
DT Journal
LA English
AB ***Sequencing*** by ***hybridization*** (SBH) is a promising alternative to the classical DNA sequencing approaches. However, the resolving power of SBH is rather low: with 64 kb sequencing chips, unknown DNA fragments only as long as 200 bp can be reconstructed in a single SBH expt. To improve the resolving power of SBH, positional SBH (PSBH) has recently been suggested; this allows (with addnl. exptl. work) approx. positions of every l-tuple in a target DNA fragment to be measured. We study the positional Eulerian path problem motivated by PSBH. The input to the positional eulerian path problem is an Eulerian graph G(V, E) in which every edge has an assocd. range of integers and the problem is to find an Eulerian path e1,...,eE in G such that the range of ei contains i. We show that the positional Eulerian path problem is NP-complete even when the max. outdegree (in-degree) of any vertex in the graph is 2. On a pos. note we present polynomial ***algorithms*** to solve a special case of PSBH (bounded PSBH), where the range of the allowed positions for any edge is bounded by a const. (it corresponds to accurate exptl. measurements of positions in PSBH). Moreover, if the positions of every l-tuple in an unknown DNA fragment of length n are measured with O(log n) error, then our ***algorithm*** runs in polynomial time. We also present an est. of the resolving power of PSBH for a more realistic case when positions are measured with .PHI.(n) error.
TI Positional ***sequencing*** by ***hybridization***
AB ***Sequencing*** by ***hybridization*** (SBH) is a promising alternative to the classical DNA sequencing approaches. However, the resolving power of SBH is rather low:. . . when the max. outdegree (in-degree) of any vertex in the graph is 2. On a pos. note we present polynomial ***algorithms*** to solve a special case of PSBH (bounded PSBH), where the range of the allowed positions for any edge is. . . positions of every l-tuple in an unknown DNA fragment of length n are measured with O(log n) error, then our ***algorithm*** runs in polynomial time. We also present an est. of the resolving power of PSBH for a more realistic case. . .

CS Center Biol. Biotechnology, Argonne National Lab., Argonne, IL, 60439-4833, USA
SO J. Comput. Biol. (1995), 2(2), 355-70
CODEN: JCOBEM; ISSN: 1066-5277
DT Journal
LA English
AB A format 1 technol. for performing massive hybridization expts. has been developed as part of the ***sequencing*** by ***hybridization*** (SBH) project. Arrays of tens of thousands of clones are interrogated with short oligomer probes to det. sets of oligomers that are present in individual clones. SBH requires highly discriminative hybridizations with a large no. of probes. One of the main uses of a reconstructed DNA sequence is in a similarity search against databases of known DNA. The authors argue that sequence reconstruction, even partial, should not be performed for this particular purpose; the authors provide and information-theoretic proof that the oligomer lists obtained from hybridization expts. should be used directly for similarity searches. The authors propose a similarity search method that takes full advantage of the subword structure of pos. identified oligomers within a clone. The method tolerates error in hybridization expts., requires fewer probes than necessary for sequencing, and is computationally efficient. To enable direct sequence recognition, the authors apply the recently developed method of sequence comparison that is based on minimal length encoding and algorithmic mutual information. The method has been tested on both real and simulated data and has led to a correct identification of clones based on hybridizations with 109 short oligomer probes. The method is applicable to hybridization data that comes from both format 1 and format 2 (***sequencing*** chip) ***hybridization*** expts. The sequence recognition method can provide targeting information for large-scale DNA sequencing by gel-based methods or by hybridization.
AB A format 1 technol. for performing massive hybridization expts. has been developed as part of the ***sequencing*** by ***hybridization*** (SBH) project. Arrays of tens of thousands of clones are interrogated with short oligomer probes to det. sets of oligomers. . . 109 short oligomer probes. The method is applicable to hybridization data that comes from both format 1 and format 2 (***sequencing*** chip) ***hybridization*** expts. The sequence recognition method can provide targeting information for large-scale DNA sequencing by gel-based methods or by hybridization.
IT ***Algorithm***
(method of sequence comparison based on minimal length encoding and algorithmic mutual information)
IT Deoxyribonucleic acid sequence determination
(***sequencing*** by ***hybridization*** (SBH); method of sequence comparison based on minimal length encoding and algorithmic mutual information)

64 ANSWER 7 OF 61 CAPLUS COPYRIGHT 1996 ACS DUPLICATE 3
AN 1995:817554 CAPLUS
DN 123:247808
TI Rconstructing strings from substrings
AU Skiena, Steven; Sundaram, Gopalakrishnan
CS Dep. Computer Sci., State Univ. of New York, Stony Brook, NY, 11794-4400, USA
SO J. Comput. Biol. (1995), 2(2), 333-53
CODEN: JCOBEM; ISSN: 1066-5277
DT Journal
LA English
AB We consider an interactive approach to DNA ***sequencing*** by ***hybridization*** , where we are permitted to ask questions of the form "is s a substring of the unknown sequence S", where s is a specific query string. We are not told where s occurs in S, nor how many times it occurs, just whether or not s a substring of S. Our goal is to det. the exact contents of S using as few queries as possible. Through interactin, far fewer queries are necessary than using conventional fixed ***sequencing*** by ***hybridization*** (SBH) sequencing chips. Wer provide tight bounds on the complexity of reconstructing unknown strings from substring queries. Our lower bound which holds even for a stronger model that returns the no. of occurrences of s as a substring of S, relies on interesting arguments based on the Bruijn sequences. We also demonstrate that subsequence queries are significantly more powerful than substring queries, matching the information theoretic lower bound. Finally, in certain applications, something may already be known about the unknown string, and hence it can be detd. faster than an arbitrary string. We show that building an optimal decision tree is NP-complete, then give an approxn. ***algorithm*** that gives tress within a const. multiplicative factor of optimal.
AB We consider an interactive approach to DNA ***sequencing*** by ***hybridization*** , where we are permitted to ask questions of the form "is s a substring of the unknown sequence S", where. . . contents of S using as few queries as possible. Through interactin, far fewer queries are necessary than using conventional fixed ***sequencing*** by ***hybridization*** (SBH) sequencing chips. Wer provide tight bounds on the complexity of reconstructing unknown strings from substring queries. Our lower bound. . . detd. faster than an arbitrary string. We show that building an optimal decision tree is NP-complete, then give an approxn. ***algorithm*** that gives tress within a const. multiplicative factor of optimal.
ST DNA sequence detn ***computer*** model ***algorithm***
IT ***Algorithm***
Computer program
Deoxyribonucleic acid sequence determination
Molecular modeling
Nucleic acid hybridization
(an interactive approach to DNA ***sequencing*** by ***hybridization*** , where questions are asked of the form "is s a substring of the unknown sequence S", where s is a specific query string)